HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

MASTERARBEIT

Herr
Christian Bausch, B.Sc.

Vorhersage der Passagierrouten im Lufttransportwesen mittels Anwendung der logistischen Regression und anderer statistischer Modelle

2017

Fakultät **Mathematik/Naturwissenschaften/Informatik**

MASTERARBEIT

Vorhersage der Passagierrouten im Lufttransportwesen mittels Anwendung der logistischen Regression und anderer statistischer Modelle

Autor:
Christian Bausch, B.Sc.

Studiengang:
Applied Mathematics

Seminargruppe:
MA13w1-M

Erstprüfer:
Professor Dr. rer. nat. Kristan Schneider

Zweitprüfer:
Dipl.-Vw. Klaus Lütjens

Mittweida, Juli 2017

I. Inhaltsverzeichnis

II. Vorwort

„Die längste Reise beginnt mit dem ersten Schritt.”
Laotse, 6. Jahrhundert vor Christus
64. Kapitel des Tao Te King

1 Einleitung

1903. Dieses Jahr gilt als das des ersten motorisierten dauerhaften Fluges, durchgeführt von den Gebrüdern Wright [AAS]. Der Beginn einer bis heute fortdauernden Ära. Bis zum Aufbau einer ersten Luftfahrtindustrie von 1909 bis zum Ende des ersten Weltkrieges vergingen lediglich 6 Jahre [HWF]. Der eigentliche Boom begann aber erst nach Ende des zweiten Weltkrieges mit Beginn der globalisierten Marktwirtschaft, als erkannt wurde, dass die stetig beschleunigte Welt schnellere Transportmöglichkeiten benötigte als Schiffe, Eisenbahnen und Automobile.
Seither wächst die Luftfahrtindustrie stärker als die meisten anderen Industrien [ICAO07]. Sie ist korreliert mit dem weltweiten ökonomischen Wachstum, aber in viel stärkeren Raten. Allein von 1975 bis 2005 betrug das jährliche durchschnittliche Wachstum der Passagierzahl bei 5.7% [ICAO07]. 2011 wurden 2.7 Milliarden Passagiere gezählt. Seit 1980 existieren pro Flughafen 80% mehr Verbindungen, die Anzahl der Abflüge stieg um 140% [LEA16]. Auch heute noch ist die Luftfahrtindustrie ein gigantischer Wachstumsmarkt. Existieren heute 42 Megastädte im Bereich des Luftverkehrs (hauptsächlich in Europa, der US-amerikanischen Ostküste, der asiatischen Südostküste und einzelne Städte im Nahen Osten wie Dubai) so wächst diese Zahl bis 2033 auf 91 an. Brasilien, der Nahe Osten und Teile Afrikas gewinnen an Bedeutung [LEA16].

Ein derart gigantisches Netzwerk ist nicht einfach zu verwalten. Schon heute kann der Ausfall eines einzigen Flughafens zu massiven globalen Ausfällen im Flugverkehr führen, weil ein einzelner Mensch „Bombe“ ruft. Oder Umweltereignisse wie der Ausbruch des Eyjafjallajökull 2010, welcher große Teile des europäischen Flugverkehrs lahmlegte und nicht nur zu massiven Verspätungen der Passagiere, sondern auch zu enormen Verlusten in Milliardenhöhe führte. Risiken birgt aber auch das ganz normale Wachstum. So sagt Reynolds [REY07] vorher, dass bei weiterem Wachstum des Luftverkehrs ohne Änderungen in den Flugrouten, der Größe der Flugzeuge, der Infrastruktur an den Flughäfen oder den Zubringemöglichkeiten die Anzahl der durchschnittlichen Verspätungen und Ausfälle unrealistisch hoch sein würden, wie auch [WAI] und [DCAA].
Ein solches Wachstum lässt schließen, dass Reaktionen auf Änderungen nicht erst passieren dürfen, wenn sie auftreten oder massiv geworden sind. Frühzeitige Kenntnisse über derlei Situationen und langfristige Planung sind der Schlüssel für ein erfolgreiches Handling potentiell in Erscheinung tretender Probleme. So zu Beispiel sieht der FLIGHTPATH 2050, Europas strategische Flugforschungsagenda, vor, dass ein Mensch innerhalb des europäischen Raumes für eine Reise vom Austritt aus der Wohnungstür bis zum Betreten der Zimmertür seines Zielortes maximal 4 Stunden benötigen soll [ACA]. Derartig schnelle Verbindungsmöglichkeiten werden bei anhaltendem Wachstum der Passagierzahlen nötig sein. Grundlage dafür bilden nicht nur jahrelang gewonnene Erfahrungen, sondern auch ausgeklügelte Prognosemodelle, die von Luftfahrttechnikern, Mathematikern und Ingenieuren aus aller Welt entwickelt wurden.

Auch die vorliegende Arbeit beschäftigt sich mit Voraussagen im Bereich des Luftverkehrs. So besteht ihr Hauptanliegen in der Erstellung eines Prognosemodells, welches zu einem Zeitpunkt in nicht allzu ferner Zukunft für eine gegebene Anzahl an Flugpassagieren bei bekannten Start- sowie Zielflughäfen ihre Verteilung auf die wichtigsten Flugrouten vorhersagt. Diese vom Deutschen Zentrum für Luft- und Raumfahrt vorgegebene Zielstellung ist mit den Mitteln der Mathematik zu bearbeiten. Dazu werden fünf verschiedene Modelle der Statistik herangezogen, erklärt, bearbeitet und verglichen. Das Augenmerk liegt dabei auf der Balance von Geschwindigkeit, Prognosegüte und Anpassung des Modells an die spezifische Problemstellung.
Ein sekundäres Ziel stellt die Aufgabe dar, wichtige, in den Hauptmodellen verwendete Daten zu den Flugrouten des Vorhersagezeitpunktes (oder eines anderen beliebigen Zeitpunktes) zu prognostizieren.
Nach der Erklärung der Aufgabenstellung in Kapitel 2 und der Angabe der wichtigsten Definitionen erfolgt in Kapitel 3 die Einführung und Spezifikation der Daten sowie ihre Aufbereitung für die Nutzung in den Modellen. Vor Beginn des eigentlichen Hauptteils gibt Kapitel 4 allgemeine Hinweise zu Verständnis, Umgang und Auswertung statistischer Modelle. Danach wird in Kapitel 5 die Theorie zu den Hauptmodellen dieser Arbeit vermittelt. In vergleichbarer Manier schließt sich mit Kapitel 6 der theoretische Hintergrund der sekundären Modelle an. Kapitel 7 führt diese Arbeit schließlich mit einer Auswertung zum Abschluss.

2 Die Aufgabenstellung

2.1 Aufgabensteller

Der Auftraggeber für diese Arbeit ist das „DLR“, das Deutsche Zentrum für Luft- und Raumfahrt. Die betreuende Person von Seiten des DLR ist die wissenschaftliche Mitarbeiterin, Frau Dipl.-Math. Katrin Kölker, der Hamburger Zweigstelle für den Bereich Lufttransportsysteme; Unterbereich Lufttransportbetrieb und –infrastrukturen.
Wie auf der hauseigenen Website [DLR] einsehbar, ist es das Forschungszentrum der Bundesrepublik Deutschland für Luft- und Raumfahrt. Die Betätigungsfelder der Forschung und Entwicklung konzentrieren sich in den Bereichen Luftfahrt, Raumfahrt, Energie, Verkehr und Sicherheit. Weiterhin obliegen dem DLR die Zuständigkeiten der deutschen Raumfahrt.

2.2 Aufgabenstellung

2.2.1 Offizielle Aufgabenstellung

Die Bewerbung der Masterarbeit lautet auf nachfolgende Ausschreibung der Helmholtz-Gesellschaft in Verbindung mit dem Deutschen Zentrum für Luft- und Raumfahrt [TSS].
*„Vorhersagen für den Luftverkehr der Zukunft basieren im Wesentlichen auf zwei Netzwerken. Zunächst werden beim **Nachfragenetzwerk** Verbindungen auf Basis von Start- und Endpunkten einer Reise eines Passagiers definiert. Das **Streckennetzwerk** hingegen erfasst die Verbindungen, auf denen Flüge angeboten werden. Der Fokus dieser Masterarbeit liegt darin, das Verständnis für die Interaktion dieser beiden Netze zu erweitern, indem die gegenseitige Abhängigkeit von Passagierströmen und Flugstreckenauslastung zwischen spezifischen Städten oder Flughäfen untersucht wird. Nach einer empirischen Analyse soll dazu ein Modell zur Abbildung dieses Zusammenhangs entwickelt und implementiert werden.*

Im Rahmen der Arbeit sollen unter anderem die folgenden Aufgaben bearbeitet werden:

1. *Einarbeitung in die Thematik und Hintergrundrecherche zu den Themen Netzwerke im Lufttransportsystem, Nachfrage- und Streckennetzwerke, Vorhersage von Netzwerken;*

2. *Analyse der Zusammenhänge zwischen Nachfrage- und Streckennetzwerken;*

3. *Erstellung eines Modells zur Ableitung eines Streckennetzwerkes aus historischen Netzwerken und vorhergesagten Nachfragenetzwerken;*

4. *Ausführliche Diskussion und Dokumentation des gewählten Ansatzes sowie der Vorgehensweise und der Ergebnisse.“*

2.2.2 Spezifizierte Aufgabenstellung

Die Speicherung großer Datenmengen in quasi Echtzeit ist heutzutage allgegenwärtig (Stichwort „Big Data“). Für Dienstleiter sind Daten zu Kauf- und Nutzerverhalten ihrer Kunden von großer Bedeutung, da sie herangezogen werden können, um Angebot, Umsatz, Gewinn etc. zu optimieren. Der korrekte Umgang mit großen Datenmengen ist oft nicht klar. Das klassische Datamining beschränkt sich oft auf automatisierte Exploration, ohne die zugrunde legenden Modelle der Mathematik zu beachten. Der Umgang mit wissenschaftlichen, modellbasierten Analysen ist weniger verbreitet und erarbeitet. Hier sollen mathematische Methoden schrittweise erarbeitet und auf Basis der gewonnenen Erkenntnisse angewandt werden, um neue Zusammenhänge und exaktere Ergebnisse zu erhalten.
In diesem Sinne ergab sich nach Absprache mit Frau Kölker, der betreuenden Stellvertreterin des DLR und Herrn Schneider, dem betreuenden Professor der Hochschule Mittweida, folgende präzisierte Aufgabenstellung:

Zu erstellen und implementieren sind mathematische Modelle die zwei Probleme lösen, welche jeweils aus monatlichen Vorhersagen bestehen. Unterschieden wird zwischen einem primären und einem sekundären Problem (siehe unten). Die Hauptaufgabe besteht in der Lösung des **Primärproblems**; das **Sekundärproblem** ist nur bei Eintreten gewisser Umstände zu lösen. Sollten diese Umstände eintreten, muss zuerst das Sekundärproblem bearbeitet werden.

Die Modelle des Primärproblems müssen bei Eingabe sogenannter **Demand-Daten** (kurz: DD) (diese entsprechen dem Nachfragenetz) für einen bestimmten Monat die **Demand-Paths-Daten** (Kurz: DPD) (diese entsprechen dem Streckennetz) dieses Monats erzeugen. Unterstützend können die **Schedule-Daten** (diese entsprechen einem Angebotsnetz) sowie Standortdaten von Flughäfen für diesen Monat als Eingabewerte herangezogen werden. Je nach Komplexität des verwendeten Modells sind diese zusätzlichen Eingabedaten erforderlich.
Die Modellparameter werden durch die Demand-, Demand-Paths-, Schedule- und Flughafendaten der vorangegangenen Monate ermittelt. Im Rahmen von Vorhersagemodellen wird dabei von „**Training**“ gesprochen. Der zeitliche Abstand zwischen Trainings- und Vorhersagedaten soll wenigstens ein Jahr betragen.

Das Sekundärproblem tritt auf, falls die zusätzlichen Eingabedaten für den vorherzu-

sagenden Monat nicht vorhanden sind. Dann sind zuerst diese Daten mittels weiterer Modelle und der Verwendung der Daten der vorangegangenen Monate zu schätzen.

2.2.3 Grundlegende Definitionen

Zunächst müssen einige grundlegende Definitionen eingeführt werden:

Definition 2.2.1 (Route)
Eine Route (genauer: Passagierroute) besteht aus einer geordneten Menge von mindestens zwei und maximal fünf Flughäfen. Der Startflughafen steht dabei an erster, der Zielflughafen an letzter Stelle. Die Umsteigeflughäfen befinden sich dazwischen. Sie sind geordnet in der Reihenfolge, in der sie angeflogen werden.

Definition 2.2.2 (Weg)
Ein Weg wird aus einem Flughafenpaar gebildet. Dabei handelt es sich um den Start- und Zielflughafen einer Route.

Bemerkung 2.2.1
Zur Wegdefinition: Zwischen Start- und Zielflughafen befindliche Umsteigeflughäfen werden bei Wegen nicht beachtet. Demzufolge können verschiedene Routen denselben Weg besitzen oder umgekehrt kann ein Weg verschiedenen Routen zuzuordnen sein.

Definition 2.2.3 (Segment)
Ein Segment wird aus zwei aufeinanderfolgenden Flughäfen einer Route gebildet.

Definition 2.2.4 (IATA-Code)
Ein IATA-Code ist ein dreistelliger Buchstabencode, der gelegentlich auch Zahlen enthalten kann. Er wird von der International Air Transport Association (IATA) an Flughäfen, Verkehrslandeplätze, wichtige Bahnhöfe, Fluglinien und Flugzeugtypen vergeben. Die Codes der Fluglinien sind zweistellig.

Beispiel 2.2.1
Jemand möchte von Berlin nach London fliegen. Dafür wählt er eine Flugverbindung, bei der er zuerst in Frankfurt und dann in Paris in eine andere Maschine umsteigen muss. Da es zum Beispiel in Berlin mehrere Flughäfen gibt, müssen die IATA-Codes zur eindeutigen Identifizierung verwendet werden.
Die IATA-Codes lauten: TGL - Berlin-Tegel, FRA - Frankfurt am Main, CDG - Paris-Charles de Gaulle, LHR - London Heathrow.
Der Weg lautet: TGL – LHR.
Seine Route lautet: TGL – FRA – CDG – LHR.
Die Segmente lauten: TGL – FRA, FRA – CDG, CDG – LHR.
Für spätere Beispiele seien die Flughäfen VIE - Wien, AMS - Amsterdam und der Bahn-

hof XIB - Bahnhof Ingersoll, RZG - Zaragoza Delicias Bus Station angegeben.
Mit diesem Beispiel lässt sich das Primärproblem folgendermaßen veranschaulichen:
Zu Beginn des Monats Januar 2016 ist aus Befragungen, Gravitationsmodellen etc. bekannt, dass 10 000 Menschen planen, den Weg TGL – LHR zu buchen (DD). Die primären Modelle ermitteln nun, dass von diesen 10 000 Passagieren bis zum Ende des Januars 3 000 die Route TGL – VIE – LHR und 7 000 die Route TGL – FRA – CDG – LHR genutzt haben werden.

Das Beispiel verdeutlicht folgende Zielstellung: Es soll die Verteilung der Passagiere auf die verfügbaren Routen geschätzt werden. Die Existenz der Routen ohne genaue Passagierzuordnung ist beim Primärproblem als bekannt vorauszusetzen.
Die Vorhersage der Existenz jeder einzelnen Route ist Bestandteil des Sekundärproblems. Die Schätzung der DPD beziehungsweise der Schedule-Daten ist nicht möglich, da diese von betriebswirtschaftlichen Gesichtspunkten der einzelnen Fluglinien, Flughäfen sowie gesetzlichen und politischen Rahmenbedingungen abhängt über die hier keine entsprechende Datenbasis vorliegt. Die Modelle des Sekundärproblems liefern höchstens eine grobe Orientierung.

3 Daten

3.1 Grundlegende Daten

In dieser Arbeit soll zwischen zeitlich veränderbaren Daten (ab hier **dynamische Daten** genannt), zum Beispiel Passagierzahlen, und zeitlich unveränderbaren Daten (ab hier **statische Daten** genannt), zum Beispiel der Position von Flughäfen oder der Beförderungskapazität von Flugzeugen, unterschieden werden.

3.1.1 Zeitabhängige Daten - Dynamische Daten

Es gibt drei Arten von Daten, die das DLR bereitstellt. Sie liegen global vor. Aufgrund der Interkontinentalpolitik ist zu vermuten, dass die Daten der amerikanischen Firmen für innerамerikanische, amerikanisch-europäische und innereuropäische Flugverbindungen exakter und vollständiger sind als beispielsweise asiatische oder russische Flugverbindungen.

- Demand-Daten: „Prep_Demand_yyyy_mm.csv“ entsprechen dem Nachfragenetz
- Demand-Paths-Daten: „Prep_DemandPaths_yyyy_mm.csv“ entsprechen dem Streckennetz
- Schedule-Daten: „scheduledata_airport_mm_yyyy.csv“ entsprechen dem Angebotsnetz

Bemerkung 3.1.1
Alle konkret angegebenen Beispiele für diese Daten sind fiktiv, falls nicht anders angegeben und dienen Anschauungszwecken.

Demand-Daten

Die Demand-Daten existieren für jeden Monat eines Jahres zurück bis zum Jahr 2002 und repräsentieren die bekannten Informationen am Anfang eines Monats. Sie bilden den Input für die Modelle und bestehen aus den IATA-Codes der Start- und Zielflughäfen der gebuchten Wege, sowie der Anzahl der Personen, die einen solchen Weg per Buchung in Anspruch nehmen wollen.

Beispiel 3.1.1
Ein Beispiel dazu findet sich in Tabelle 3.1.

Origin	Destination	Passengers
VIE	AMS	7
TGL	LHR	345

Tabelle 3.1: Auszug aus den DD: Zwei Wege mit ihren Passagierzahlen - Prep_Demand_2014_4.csv

Dabei handelt es sich um passagierbezogene Daten, sogenannte MIDT-Daten (Marketing Information Data Tapes). MIDT-Daten enthalten detaillierte Informationen über die allgemeinen Buchungstätigkeiten von Reisebüros, Buchungsportalen wie Expedia und Luftfahrtunternehmen (nach [SAS]). Durch die Analyse der MIDT-Daten können Fluggesellschaften wichtige Informationen für Entscheidungsprozesse in verschiedenen Bereichen des Unternehmens erhalten, wie zum Beispiel Marketing und Vertrieb, Ertragsmanagement, Streckenplanung und Flugplanerstellung (nach [AIR]). Die hier verwendeten Daten werden von der Firma SABRE gesammelt, um alle Passagiere zu erfassen. SABRE ist ein führender US-amerikanischer Softwaretechnologieanbieter für die Reiseindustrie. Weitere Informationen können der firmeneigenen Website entnommen werden: [SAB].

Bemerkung 3.1.2
Es sei an dieser Stelle ausdrücklich hervorgehoben, das die Anzahl dieser Buchungen selbst lediglich aus Schätzungen stammt und beschreibt, wie viele Leute planen einen Flug zu buchen beziehungsweise planen zu fliegen.

Demand-Paths-Daten

Die Demand-Paths-Daten existieren, wie die DD, für jeden Monat eines Jahres zurück bis in das Jahr 2002. Wie die untenstehende Tabelle 3.2 zeigt, beinhalten sie die Routen, die in einem bestimmten Monat geflogen werden und die Gesamtzahl der Passagiere, die diese Routen über den ganzen Monat hinweg in Anspruch genommen haben. Dabei sollen die verfügbaren Routen am Anfang des Monats bekannt sein und die Anzahl der Passagiere erst am Ende. Für die Modelle bedeutet dies, dass die Routen Teil der Eingabedaten (dem **Input**) und die Passagiere Teil der Ausgabedaten (dem **Output**) sind. Die Darstellung der Flughäfen der Route erfolgt über ihre IATA-Codes.
Mittels der DPD werden zusätzliche Informationen über die Flughäfen gewonnen, so zum Beispiel die Wichtigkeiten der Flughäfen. Diese stehen in direkter Abhängigkeit zu den Passagierströmen und schwanken daher im monatlichen Vergleich.
Die verfügbaren DPD stammen ebenfalls von der Firma SABRE. Sowohl die DD als auch die DPD sind für die Zeit nach 2009 von der Firma SABRE bereinigt und aufbereitet. Für die Zeit davor sind die Datenbanken nicht unbedingt konsistent.

Origin	Destination	Connect Point1	Connect Point2	Connect Point3	Passengers
TGL	VIE	CDG	AMS	LHR	27
VIE	CDG	LHR	TGL		345

Tabelle 3.2: Auszug aus den Demand-Paths-Daten: Zwei Routen mit ihren Passagierzahlen - Prep_DemandPaths_2014_4.csv

Sowohl bei den DD als auch bei den DPD können IATA-Codes von Bahnhöfen auftauchen. Der Umstand ist dem Angebot von Rail-and-Fly-Tickets zu verdanken. Dieses Angebot ermöglicht einem Passagier den Kauf eines Flugtickets, bei dem die Zubringerfahrt per Bahn zusätzlich enthalten ist. In einem solchen Fall treten eventuelle Zubringerbahnhöfe am Anfang und Ende der Reise in den Daten auf. Die Bahnhöfe vor dem ersten und nach dem letzten Flughafen werden vor der Verwendung der DPD entfernt. Der in den DD entstandene Passagierunterschied wird korrigiert. DD die Bahnhöfe enthalten, werden gelöscht. Viele dieser Bahnhöfe sind allerdings mit den oft sehr nahegelegenen Flughäfen gut über die Infrastruktur angebunden und besitzen deshalb häufig denselben Code wie der Flughafen, weshalb sie nicht immer identifiziert werden.

Schedule-Daten

Die Schedule-Daten existieren ebenfalls für jeden Monat eines Jahres zurück bis zum Jahr 2002 und bilden einen unterstützenden Input zu den DD. Zu jedem Segment jeder Route der DPD desselben Monats existieren ein oder mehrere Einträge. Wie in Beispiel 3.1.2 zu sehen ist, bestehen die Einträge aus dem Start- und Zielflughafen eines Segments. Dazu vermerkt sind sowohl die Flugzeugtypen, die auf diesem Abschnitt unterwegs sind, als auch ihre geplante Anzahl an Einsätzen (Frequenz). Weiterhin enthalten sind die flugzeugtypbedingte Reisezeit und den Abstand der Flughäfen. Aufgrund unterschiedlicher Flugzeuge und Einsatzzahlen sind mehrere Eintragungen zu denselben Segmenten üblich. Für die praktische Anwendung werden der maximale und minimale Flugzeugtyp gesucht, sowie die durchschnittliche Flugzeit und Frequenz gebildet.

Beispiel 3.1.2

Origin	Destination	AcType	YearlyFrequency	FlightTime [min]	Distance [km]
TGL	LHR	Airbus A315	34	75.0	651.093
VIE	CDG	Ju-52	23	97.0	1 391.273

Tabelle 3.3: Auszug aus den Schedule-Daten - scheduledata_airport_2014_04.csv

Die Daten werden von der Firma Innovata erfasst und stellen die geplanten Flüge der Fluggesellschaften ohne Verspätung dar. Die Daten sind von besserer Qualität als die Daten des US-amerikanischen Bureau of Transportation Statistics (die sogenannten T-100 Daten) und dazu global vorhanden.
Die US-amerikanische Firma Innovata ist Teil von Flightglobal, einer Online-Nachrichten- und Informationswebsite mit den Schwerpunkten Luftverkehr und Luftfahrt. Sie ist strategischer Partner der IATA und der führende Anbieter der Flugplandaten von Fluglinien. Die Flugplandaten beziehen sich auf mehr als 800 Flugzeuge weltweit, siehe: [INN].

3.1.2 Zeitunabhängige Daten - Statische Daten

Es gibt drei Arten von Daten, welche wie die dynamischen Daten global sind. Analog zu den Schedule-Daten fungieren sie als unterstützender Input für das Training der Modelle.

- Flugzeug-Daten: „AircraftInformations.csv“
- Bahnhofs-Daten: „Railstations.csv“
- Flughafen-Daten: „AirportInformations.csv“

Flugzeug-Daten

Die Herkunft der Flugzeugdaten unterliegt in allen Fällen der eigenen Onlinerecherche, die sich zumeist auf Artikel der Internetdatenbank Wikipedia stützt. Für jede Maschine die in den Jahren 2002 – 2015 in den Scheduledaten auftrat, wurde eine Suche nach 7 Merkmalen durchgeführt:

1. **Name**: Suchbegriff bei der Onlinerecherche, bei mehreren verschiedenen Subklassen wurde versucht, die Bezeichnung der Oberklasse zu wählen und bei den nachfolgenden Eigenschaften den Durchschnitt beziehungsweise das Minimum oder Maximum zu verwenden, insofern diese nicht eine signifikante Änderung der Flugzeugcharakterisierungsnummer zur Folge haben. Interessanterweise gab es einige ungewöhnliche Flugzeuge. So ist zum Beispiel unter der Bezeichnung CV6 der amerikanische Flugzeugträger USS Enterprise zu finden.

2. **Flugzeugcharakterisierungsnummer**: Abhängig von der Passagierzahl x und Antriebstyp wird das Flugzeug in eine von 16 Klassen eingeteilt:

   - Typ 0: Oberflächenausrüstung (Autos, Schiffe, ...)
   - Typ 1: Helikopter
   - Typ 2: Wasserflugzeug
   - Typ 3: Propeller: $x \leq 20$ Passagiere
   - Typ 4: Propeller: $x > 20$ Passagiere

- Typ 5: Privatjet
- Typ 6: Düse: $x \leq 20$ Passagiere
- Typ 7: Düse: $20 < x \leq 50$ Passagiere
- Typ 8: Düse: $50 < x \leq 100$ Passagiere
- Typ 9: Düse: $100 < x \leq 150$ Passagiere
- Typ 10: Düse: $150 < x \leq 200$ Passagiere
- Typ 11: Düse: $200 < x \leq 250$ Passagiere
- Typ 12: Düse: $250 < x \leq 300$ Passagiere
- Typ 13: Düse: $300 < x \leq 400$ Passagiere
- Typ 14: Düse: $400 < x \leq 500$ Passagiere
- Typ 15: Düse: $x > 500$ Passagiere

Die Charakterisierungsnummer ist, neben dem Namen als Kennzeichnung, die einzige Kennzahl, die in die Modelle einfließt. Die restlichen Kennzahlen sind lediglich interessehalber aufgenommen, für den Fall, dass im Verlauf der Arbeit Sachverhalte anhand der Flugzeuge nachvollzogen werden müssen.

3. Beim **Antriebstypen** wird grundsätzlich in die 5 Kategorien Helikopter, Wasserflugzeug, Propeller, Privatjet und Düsen unterteilt. Chimären aus verschiedenen Typen wie Amphibienflugzeuge wurden subjektiv mit dem passendsten Typen bewertet.

4. Die **Passagierzahl** richtet sich meist nach dem Maximum und dem Minimum der durchschnittlich vorhandenen Sitzplätze der verschiedenen Subklassen. Für die Charakterisierungsnummer ging die am häufigsten auftretende Sitzplatzanzahl ein.

5. Für die **Geschwindigkeit** wurde die durchschnittliche Reisegeschwindigkeit gewählt. Mit den Werten der Subklassen wurde analog zur Passagierzahl verfahren.

6. Der **Reichweitentyp** beschreibt die Art der Flugstrecken auf denen ein Flugzeug üblicherweise eingesetzt wird. Nach der Fluggastrechteverordnung der EU wird in Kurzstrecken bis 1 500 km, Mittelstrecken von 1 500 km bis 3 500 km und Langstrecken ab 3 500 km eingeteilt.

7. **Die Reichweite** gibt die durchschnittliche Reichweite bei Normalbelastung und unter Normbedingungen an.

Beispiel 3.1.3
Tabelle 3.4 illustriert den Inhalt der Flugzeug-Daten

Aircraft Name	Chr.Nr.	Mach.Type	Passengers	Speed in km/h	Range Type	Range in km
BOEING 767	13	Düse	255-375	851	long	12 223
Airbus A340	13	Düse	261-419	890	long	16 700

Tabelle 3.4: Auszug aus AircraftInformations.csv

Bahnhof-Daten

Die Bahnhof-Daten werden aus einer Liste mit IATA-Codes gebildet. Diese Liste ist die Sammlung aller Bahnhöfe, Häfen oder Busstationen, die während der Datenrecherche aufgetreten sind. In einigen Fällen war die Existenz eines Flughafens nicht nachweisbar. Aufgrund der geographischen Gegebenheiten konnte oftmals davon ausgegangen werden, dass es sich bei dem IATA-Code um einen Bahnhof oder Hafen handelt.

Beispiel 3.1.4
Tabelle 3.5 illustriert den Inhalt der Bahnhof-Daten

Railstations IATA/ICAO Codes
XIB
RZG

Tabelle 3.5: Auszug aus Railstations.csv

Airport-Daten

Die Airport-Daten wurden teilweise vom DLR gestellt beziehungsweise in Eigenrecherche ermittelt.
Seiten der manuellen Recherche:

- [LET] https://www.3lettercode.de
- [APB] http://airportsbase.org
- [DIF] http://www.distancesfrom.com
- [FLS] http://www.flightstats.com
- [OFS] http://www.openflights.org/
- [PRO] http://www.prokerala.com/travel/airports/
- [TWM] https://tools.wmflabs.org/geohack/
- [WAC] http://www.world-airport-codes.com/

Dabei handelt es sich um die standortbezogenen Daten der Flughäfen, die keinen zeitlichen Änderungen unterliegen und somit nur einmalig vorhanden sind. Die Datei „AirportInformations.csv“ ist eine Zusammenfassung aus mehreren Einzeldateien, die im Verlaufe der Arbeit gestellt beziehungsweise erstellt wurden. Für jeden Flughafen wurde versucht, folgende Kenndaten aufzunehmen. Sollte eine Information nicht vorhanden sein, steht ein leerer Eintrag an der entsprechenden Stelle.

1. **IATA-Code**: Siehe Definition 2.2.4 auf Seite 5.

2. **ICAO-Code**: Ist ein vierstelliger Buchstabencode, der in seltenen Fällen auch Zahlen enthalten kann und den IATA-Codes entspricht. Sie werden von der Internationalen Zivilen Luftverkehrsorganisation („International Civil Aviation Organization“) vergeben. Er dient programmintern zur Identifizierung eines Flughafens, sollte der IATA-Code zum Bearbeitungszeitpunkt noch nicht vorhanden sein.

3. **Flughafenname**

4. **Stadt des Flughafens**

5. **Region des Landes**

6. **Land**

7. **Kontinent**: Kann aufgrund der Größe eines Kontinents differenziert sein (zum Beispiel Nord-, Süd- und Mittelamerika).

8. **Latitude**: Beschreibt die geographische Breite im Intervall [-90°, 90°] (Süd-Nord-Ausbreitung).

9. **Longitude**: Beschreibt die geographische Länge im Intervall [-180°, 180°] (West-Ost-Ausdehnung).

10. **Altitude**: Beschreibt die Höhe des Flughafens über dem Meeresspiegel.

11. **Zeitzone** des Flughafens.

12. **DST**: Daylight Saving Time, also Sommer- oder Winterzeit.

13. **Städtische Population**

14. **GDP** der Stadt: Steht für Gross Domestic Product (Bruttoinlandsprodukt) und beschreibt für ein Quartal oder ein Jahr die ökonomische Wirtschaftskraft eines Landes oder Region.

Die Nummern 3, 4, 5, 10, 11, 12, 13, 14 gehen nicht in die Berechnungen mit ein und sind nur interessehalber mit aufgeführt. Der Hauptgrund dafür liegt in der Unvollständigkeit der Daten. So ist die Altitude lediglich bei 10% der Flughäfen vorhanden.

Wie bereits erwähnt, wurden einige Daten gestellt beziehungsweise in Eigenrecherche erstellt. Die Daten werden im Folgenden nur mit Namen und einem beispielhaftem Auszug vorgestellt.

Beispiel 3.1.5
Das Beispiel für die Airportdaten ist in Tabelle 3.7 aufgeführt.

Tabelle 3.6 beinhaltet die Zuordnungen der anschließend aufgeführten Tabellen zu ihren Namen, Quellen und Referenzen.

Name	Quelle	Tabellen-referenz
AirportDataManual.csv	Vom DLR gestellt und recherchiert, Recherchequellen siehe oben	Tabelle 3.8 Seite 16
Airports_OpenFlights.csv	Vom DLR gestellt beziehungsweise recherchiert, Recherchequellen siehe oben	Tabelle 3.9 Seite 16
AirportsDatabase1.0.csv	Vom DLR gestellt	Tabelle 3.10 Seite 17
GlobalAirportDatabase.xlsx	[PAR]	Tabelle 3.11 Seite 17
IATA_Airline_Codes.xlsx	[IAC]	Tabelle 3.12 Seite 17
IATA_Airport_Codes.xlsx	[IFC]	Tabelle 3.13 Seite 18
IATA_Railstation_Codes.xlsx	[IBC]	Tabelle 3.14 Seite 18
LandNamenDeutschEnglisch.xlsx	Recherchiert, Recherchequellen siehe oben	Tabelle 3.15 Seite 18
self_found_additional_IATA_Codes.csv	Recherchiert, Recherchequellen siehe oben	Tabelle 3.16 Seite 18
self_found_again_additional_IATA_Codes.csv	Recherchiert, Recherchequellen siehe oben	Tabelle 3.17 Seite 19
self_found_IATA_Codes.csv	Recherchiert, Recherchequellen siehe oben	Tabelle 3.18 Seite 19
self_found_Railway_Stations.csv	Recherchiert, Recherchequellen siehe oben	Tabelle 3.19 Seite 19
StädteGDPPopu.xlsx	Vom DLR gestellt	Tabelle 3.20 Seite 19

Tabelle 3.6: Gibt Tabellennamen mit ihren Quellen und Referenznummern an

IATA	ICAO	Airport	Town	Region	Country	Continent	Lat.	Long.	Alt.	Tm.zn.	DST	Pop.	GDP
AAY	OYGD	Al Ghaidah Intl	Al Ghaydah	Al-Mahrah	Yemen	Middle East	16.2	52.2	41	3	U	10948	2.3E+7
AAE	DABB	Annaba Airport	Annaba	Annaba	Algeria	Africa	36.8	7.8	5	1	N	206570	1.5E+9

Tabelle 3.7: AirportInformations.csv

iata_code	latitude_deg	longitude_deg	SOURCE
AAA	-17.35	-145.51	OurAirports.com – Database
AAM	-24.82	31.54	OurAirports.com - Database

Tabelle 3.8: AirportDataManual.csv

Apt ID	Name	City	Country	IATA/FAA	ICAO	Lat.	Long.	Alt. (ft)	Tm.zn.	DST	Continent
1	Goroka	Goroka	Papua New Guinea	GKA	AYGA	-6.082	145.39	5282	10	U	Pacific/Port_Moresby
15	Isafjordur	Isafjordur	Iceland	IFJ	BIIS	66.06	-23.14	8	0	N	Atlantic/Reykjavik

Tabelle 3.9: Airports_OpenFlights.csv

LY Apt. ID	IATA	Name	ADICity	ADICountry	ADIRegion	ICAO	Latitude Apt.	Longitude Apt.	Altitude (ft)
1	04G	Lansdowne Airport	Youngstown	USA			41.13	-80.62	1044
6390	SFA	Thyna	Sfax	Tunisia	Africa	DTTX	34.72	10.69	85

Timezone	DST	City IATA	City Latitude	City Longitude	Num_CityID	Non Apt. (eg Railway, Busstop)	Source	Comments
-5	A		41.1	-80.65			OpenFlights	
1	E	SFA	34.74	10.76	2939		OpenFlights	

Tabelle 3.10: AirportsDatabase1.0.csv

ICAO	IATA	Apt.	Town	Country	Lat.Deg.	Lat.Min.	Lat.Sec.	Lat.Dir.	Long.Deg.	Long.Min.	Long.Sec.	Long.Dir.	Alt.
LTBL	IGL	CIGLI	IZMIR	Turkey	38	30	47	N	27	0	36	E	17
MHTE	TEA	TELA	TELA	Honduras	15	46	33	N	87	28	32	U	7

Tabelle 3.11: GlobalAirportDatabase.xlsx

Code	Fluggesellschaft	Land	Bemerkung
0D	Darwin Airline	Schweiz	neuer IATA-Code ist F7
V7	Volotea	Spanien	vorher Air Senegal

Tabelle 3.12: IATA_Airline_Codes.xlsx

IATA	ICAO	Flughafen	Ort	Region	Land
AAA	NTGA	Flughafen Anaa	Anaa	Tuamotu-Archipel	Französisch-Polynesien
QOW	DN52	Flughafen Owerri	Owerri	Imo	Nigeria

Tabelle 3.13: IATA_Airport_Codes.xlsx

IATA Code	Bahnhof	Stadt	Land
SPL	Bahnhof Schiphol	Amsterdam	Niederlande
XWL	Göteborgs centralstation	Göteborg	Schweden

Tabelle 3.14: IATA_Railstation_Codes.xlsx

LandNameEnglisch	LandNameDeutsch
Albania	Albanien
Lithuania	Litauen

Tabelle 3.15: LandNamenDeutschEnglisch.xlsx

IATA Code	Latitude	Longitude	Name
ZBV	39.633811	-106.509833	Beaver Creek Van Service
MTP	41.0733	-71.9235	Montauk Airport

Tabelle 3.16: self_found_additional_IATA_Codes.csv

IATA-Code	ICAO-Code	Airport	Town	Region	Country	Latitude	Longitude
KVN		Kunshan South Railway Station	Kunshan	Jiangsu	China	31.355	120.9464
POV	LZPW	Presov Airport	Presov		Slovakia	49.03	21.32

Tabelle 3.17: self_found_again_additional_IATA_Codes.csv

IATA Code	Airport Name	Town	Region	Country	ICAO Code	Latitude	Longitude
TXF	Teixeira de Freitas Airport	Teixeira de Freitas		Brazil		-17.5245	-39.6685
JWN	Zanjan Airport	Zanjan		Iran	OITZ	36.774418	48.369854

Tabelle 3.18: self_found_IATA_Codes.csv

Railway Stations
XEO
ZAS

Tabelle 3.19: self_found_Railway_Stations.csv

IATA	City	Country	Region	Lat	Long	Population	GDP
HND	Tokyo	Japan	Asia	0,622899	2,43808	8336599	2,63E+11
DLC	Dalian	China	Asia	0,679146	2,12236	2035307	1,6E+10

Tabelle 3.20: StädteGDPPopu.xlsx

3.2 Datenvorbereitung

Die Daten für die Modelle der Kapitel 5 und 6 müssen bereinigt werden. Wie schon in Abschnitt 3.1 erwähnt, sind zum Beispiel Bahnhöfe enthalten oder Routen sind unwichtig (beispielsweise weniger als fünf Personen nutzen diese Route), was sie entbehrlich macht. Ebenso können die Daten zusammengefasst werden um später neue Erkenntnisse aus ihnen zu gewinnen.
Alle Datensätze die ein Datum enthalten, erhalten zuerst einen neuen Namen des folgenden Typs: „Name_yyyy_mm .Dateityp“ .

3.2.1 Bereinigung der dynamischen Daten

Bei den zu bereinigenden dynamischen Daten handelt es sich lediglich um die DD und die DPD. Die Schedule-Daten sind nicht zu überprüfen, da sie zusammen mit den stationären Daten unterstützenden Input bilden und nur für diejenigen Routen und Wege von Bedeutung sind, welche die Bereinigung überstanden haben. Zur Vermeidung sich wiederholender Aussagen sei an dieser Stelle gesagt, dass bei den nachfolgend beschriebenen Arbeitsschritten am Ende sämtliche Wege und Routen einer Datei auf eine multiple Existenz hin überprüft werden. Sollte dies der Fall sein, so werden die betroffenen Passagiere summiert und nur eine Weg-/Routeneintragung mit der Passagiersumme zurückbehalten.

Beispiel 3.2.1
In einer Datei befinden sich auf mehrere Zeilen verteilte Routen.

LHR	TGL	VIE	CDG	AMS	234
TGL	CDG	VIE	AMS	LHR	56
LHR	TGL	VIE	CDG	AMS	37
TGL	CDG	VIE	AMS	LHR	4

Die erste und dritte, sowie die zweite und vierte Zeile beschrieben dieselben Routen. Die getrennten Eintragungen werden nun zusammengefasst.

LHR	TGL	VIE	CDG	AMS	271
TGL	CDG	VIE	AMS	LHR	60

1. Arbeitsschritte für gemeinsame Demand- und Demand-Paths-Daten

Zuerst werden DD und DPD aufgerufen, die zum selben Zeitpunkt existieren, insofern sie vorliegen. Es kann während der Arbeitsschritte der Fall eintreten, dass Routen verändert werden, sodass sie einem anderen Weg zuzuordnen sind. Dann ist die Anzahl

der Routenpassagiere in der zeitlich zuzuordnenden Demand-Datei vom alten Weg zu subtrahieren und zum neuen Weg zu addieren. Bei Löschung einer Route sind lediglich die Passagiere vom entsprechenden Weg zu subtrahieren. Dieser Vorgang wird durch untenstehendes Beispiel 3.2.2 verdeutlicht. Folgende Schritte sind durchzuführen:

1. Entferne alle Wege, die einen Bahnhof enthalten ($\approx$ 4 200 - 8 000 betroffene Wege).

2. Entferne alle Wege, die zu sich selbst zurückführen (bisher kein Weg).

3. Entferne die Bahnhöfe in den DPD vor dem ersten und nach dem letzten Flughafen. Diese Zubringerbahnhöfe sind nicht von Interesse. Die Bahnhöfe dazwischen werden behalten, um nicht vollständige Routen zu zerteilen. Die Anzahl der betroffenen Routen ist ohnehin gering ($\approx$ 5 000 - 12 000 betroffene Routen, was weniger als einem Prozent entspricht).

4. Entferne einen Flughafen, wenn er direkt danach nocheinmal erscheint. Aus LHR - TGL - TGL - CDG mache LHR - TGL - CDG. Dabei handelt es sich vermutlich um irrelevante flughafeninterne Zubringerfahrten per Bus oder Bahn ($\approx$ 3 Routen für die Zeit ab April 2010, vorher durchschnittlich 200 Routen was vermutlich mit der veränderten Quelle der Datensätze zusammenhängt. Eine andere Erklärung ist, dass es sich dabei um einen Flug mit Übernachtung handelt. Sollte nur ein Flughafen übrig bleiben, so wird er entfernt.

5. Entfernen von Routen mit zu wenigen Passagieren mittels eines Schwellwertes. Weitere Erläuterungen dazu siehe Abschnitt 3.2.1. Sollte es Wege geben, die dadurch nur noch eine Route beinhalten, so wird eine weitere Route hinzugefügt, deren Passagierzahl aber 70% des vorher verwendeten Grenzwertes übersteigt. Dadurch bleiben $10-15\%$ der Routen übrig, die 95% der Passagiere beinhalten. Der Anteil der hinzugefügten Routen niederer Passagierzahl liegt bei ungefähr 1% der übriggebliebenen Routen höherer Passagierzahl.

6. Haben Routen keinen Eintrag als Wege in den DD, wird der fehlende Wegeintrag erzeugt (bisher ist kein derartiger Fall aufgetreten).

7. Entferne Wege, die keine Routen in den DPD besitzen (bisher ist kein derartiger Fall aufgetreten).

Beispiel 3.2.2

Die nachfolgenden Tabellen 3.21 - 3.24 illustrieren die beschriebenen Arbeitsschritte.

Start	Flughafen 1	Flughafen 2	Flughafen 3	Ziel	Passagiere
LHR	LHR	CDG	AMS	TGL	30
LHR	CDG	CDG	AMS	TGL	40
LHR	CDG	AMS		TGL	120
VIE				TGL	30
RZG	VIE			TGL	40

Tabelle 3.21: Verschiedene Routen mit doppelten Einträgen und einer Route mit Bahnhof RZG - Zustand vor Durchführung der Arbeitsschritte

Start	Ziel	Passagiere
LHR	TGL	190
VIE	TGL	30
RZG	TGL	40

Tabelle 3.22: Wege zu den oben aufgeführten Routen - Zustand vor Durchführung der Arbeitsschritte

Start	Flughafen 1	Flughafen 2	Flughafen 3	Ziel	Passagiere
LHR	CDG	AMS		TGL	190
VIE				TGL	70

Tabelle 3.23: Routen nach Durchführung der Arbeitsschritte: Doppelte Flughäfen wurden entfernt, ebenso Bahnhöfe am Beginn und mehrfache Routen zusammengefasst.

Start	Ziel	Passagiere
LHR	TGL	190
VIE	TGL	70

Tabelle 3.24: Wege nach Durchführung der Arbeitsschritte: Der Weg RZG - TGL wurde entfernt; da aus der Route RZG - VIE - TGL die Route VIE - TGL wurde, mussten die 40 betroffenen Passagiere umgeschrieben werden, sodass die Passagieranzahl beim Weg VIE - TGL korrekt ist.

Bemerkung zum Beispiel:
Sollte ein Weg dadurch keine oder eine negative Anzahl an Passagieren besitzen, so sind er und seine Routen nach Beendigung aller Arbeitsschritte zu entfernen. Die Güte der Daten zeichnet sich dadurch aus, dass durch das unten beschriebene ***Perzentil*** *jeden Monat zwischen 300 000 und 500 000 Wege mit Passagierzahl 0 entfernt werden, es bei allen vorhandenen Daten allerdings nur 6 Wege mit negativer Passagierzahl gibt.*

Somit ist lediglich ein verschwindend geringer Anteil der Routen beziehungsweise Wege fehlerhaft. Sollten Passagiere zu einem nicht existenten Weg addiert werden, so ist eine neue Eintragung vorzunehmen.

2. Arbeitsschritte für Demand-Daten ohne zugehörige Demand-Paths-Daten

Dieser Fall kann beispielsweise bei den zu schätzenden Monaten auftreten, bei denen davon auszugehen ist, das keine DPD vorliegen. Hier entfällt die Korrektur der Passagierzahl der Wege mit veränderten Routen, wie am Anfang des ersten Arbeitsschrittes beschrieben. Es betrifft lediglich die Tabellen, die Wege beinhalten. Folgende Schritte sind durchzuführen:

1. Entferne Wege, die einen Bahnhof enthalten,

2. Entferne Wege, die zu sich selbst zurückführen.

3. Arbeitsschritte für Demand-Paths-Daten ohne zugehörige Demand-Daten

Bei diesem Fall existieren zu einem bestimmten Zeitpunkt lediglich die DPD, aber keine DD, womit die Korrektur der Passagierzahl der Wege mit veränderten Routen entfällt. Die Ausgangssituation ist wie in Beispiel 3.2.2 ohne Tabelle 3.22, da diese nicht existent ist. Tabelle 3.24 wird aus der finalen Routentabelle 3.23 erzeugt. Folgende Schritte sind durchzuführen:

1. Entferne die Bahnhöfe in den DPD vor dem ersten und nach dem letzten Flughafen.

2. Entferne einen Flughafen, wenn er direkt danach nocheinmal erscheint.

3. Entferne Routen mit zu wenigen Passagieren mittels eines Schwellwertes. Weitere Erläuterungen siehe Abschnitt 3.2.1. Sollte es Wege geben, die dadurch nur noch eine Route beinhalten, so wird eine weitere Route hinzugefügt, deren Passagierzahl aber 70% des vorher verwendeten Grenzwertes übersteigt.

4. Erzeuge die zugehörige Demand-Datei. Bilde dazu aus jeder Route einen Weg und füge ihn mitsamt den Routenpassagieren in die Datei ein. Ist der Weg bereits vorhanden, addiere lediglich die Passagiere zum bereits eingefügten Weg hinzu.

Routen mit zu wenigen Passagieren - Perzentile

Im vorherigen Abschnitt 3.2.1 wurde von einer zu geringen Anzahl von Passagieren gesprochen. Dass eine Route, die lediglich einen Passagier aufweist, im Vergleich zu einer Route mit 20 000 Passagieren eine geringe Bedeutung aufweist, erscheint einleuchtend. Derart gering frequentierte Routen bergen als sogenannte „**Ausreißer**“ in vielen Modellen sogar Fehlerpotential.

Beispiel 3.2.3
Den etwa 300 000 Routen, die in einem Monat von lediglich einem Passagier in Anspruch genommen werden, stehen rund 200 Routen mit je 20 000 Passagieren pro Monat gegenüber, die also 4 Millionen Passagiere beinhalten. Wenn ein Modell bei den 200 Routen um 5 Passagiere abweicht, ist der resultierende Fehler verschwindend gering. Eine Abweichung von 5 Passagieren auf den 300 000 Einerrouten hingegen erzeugt in der Analyse einen 300 000-fachen Fehler von 500%, der den geringen Fehlerwert der 200 Routen deutlich übersteigt. Die Folge ist eine fehlerhafte Interpretation der Modellgüte, verursacht durch kleine Abweichungen der Passagierzahlen bei Routen mit ohnehin geringen Passagierzahlen.

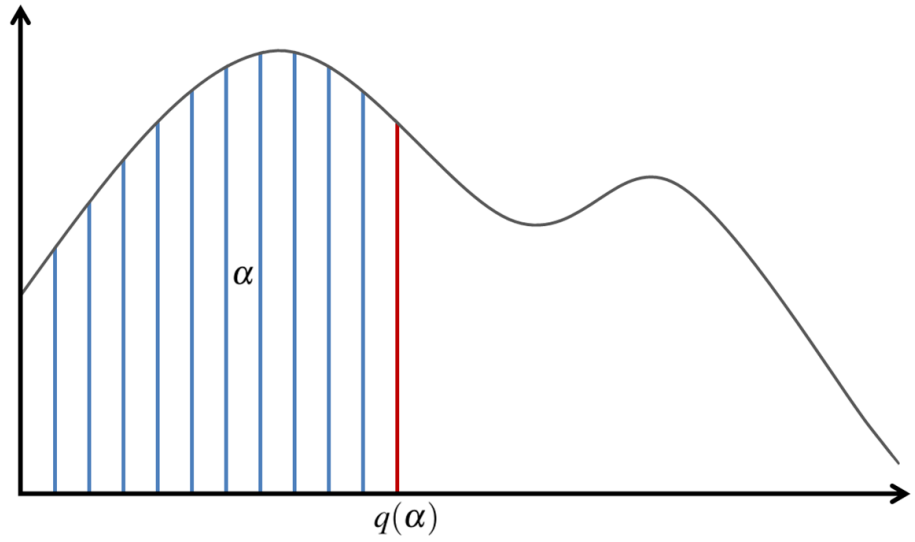


Abbildung 3.1: Eine beliebige Dichtefunktion $f(x)$ bei der sich linkseitig von $q(\alpha)$ $\alpha\%$ der Daten befinden.

Die Mathematik beseitigt derartige Ausreißer durch die Anwendung von Schwellwerten, sogenannten **Perzentilen**, siehe Definition 3.2.1. Üblich sind 1%, 2.5%, 5%, 10%, 25% und 50%-Perzentile.

Definition 3.2.1
Ein Perzentil ist ein Lagemaß in der Statistik. Gegeben sei die Zufallsvariable X mit einer Verteilungsfunktion F. Mit q wird für ein $\alpha \in (0,1)$ die Menge aller α-Perzentile von X beschrieben durch

$$q(\alpha) = \inf\{x \in \mathbb{R} | P(X \leq x) \geq \alpha\}. \tag{3.1}$$

Anschaulich gesprochen ist ein Perzentil eine Schranke die besagt, dass sich in der Dichtefunktion $f(x)$ einer Verteilungsfunktion $F(x)$ linksseitig des Schwellwertes $q(\alpha)$ $\alpha\%$ der Daten befinden.

Betrachtet sei ein Weg. Für die vorliegenden Daten sei $\alpha = 5\%$, $X =$ *Anzahl der monatlichen Passagiere einer Route des Weges* und $N =$ *Anzahl der monatlichen Passagiere des Weges*. Sei $F(X) = \frac{X}{N}$. Der Schwellwert ermittelt sich durch

$$q(\alpha) = \sup\{x \in \mathbb{R} | F(x) \leq \alpha\}. \tag{3.2}$$

Bei den vorliegenden Daten ergibt sich je nach betrachteten Zeitpunkt ein Schwellwert zwischen 40 und 100.

Das 5%-Perzentil ist ein Wert, der sich in der Mathematik im Laufe der Zeit etabliert hat. Viele kleine Routen lagen oft dicht unterhalb des Schwellwertes von 5% der Wegpassagierzahl, wohingegen die größeren Routen oft recht eindeutig über diesem Wert lagen. Auffällig ist der hohe Anteil größerer Routen nach 2010. Für weitere Informationen zur allgemeinen Flugentwicklung sei an dieser Stelle [EVA09], [WAN08], [REG11], [FAA08], [REY98], [HOL04], [BON08] und [WEI05] empfohlen.
Bei der Anwendung auf die Daten tritt der angenehme Effekt auf, dass $10 - 15\%$ der Routen übrig bleiben, die jedoch $90 - 92\%$ der Passagiere abdecken, wie bei den Abbildungen 3.2 und 3.3 zu sehen ist. Vor allem bei Abbildung 3.3 ist sehr gut erkennbar, wie das 5%-Perzentil eine recht genaue Grenze zwischen den Routen mit niedrigem Passagieraufkommen und Routen mit hohen Passagieraufkommen zieht.

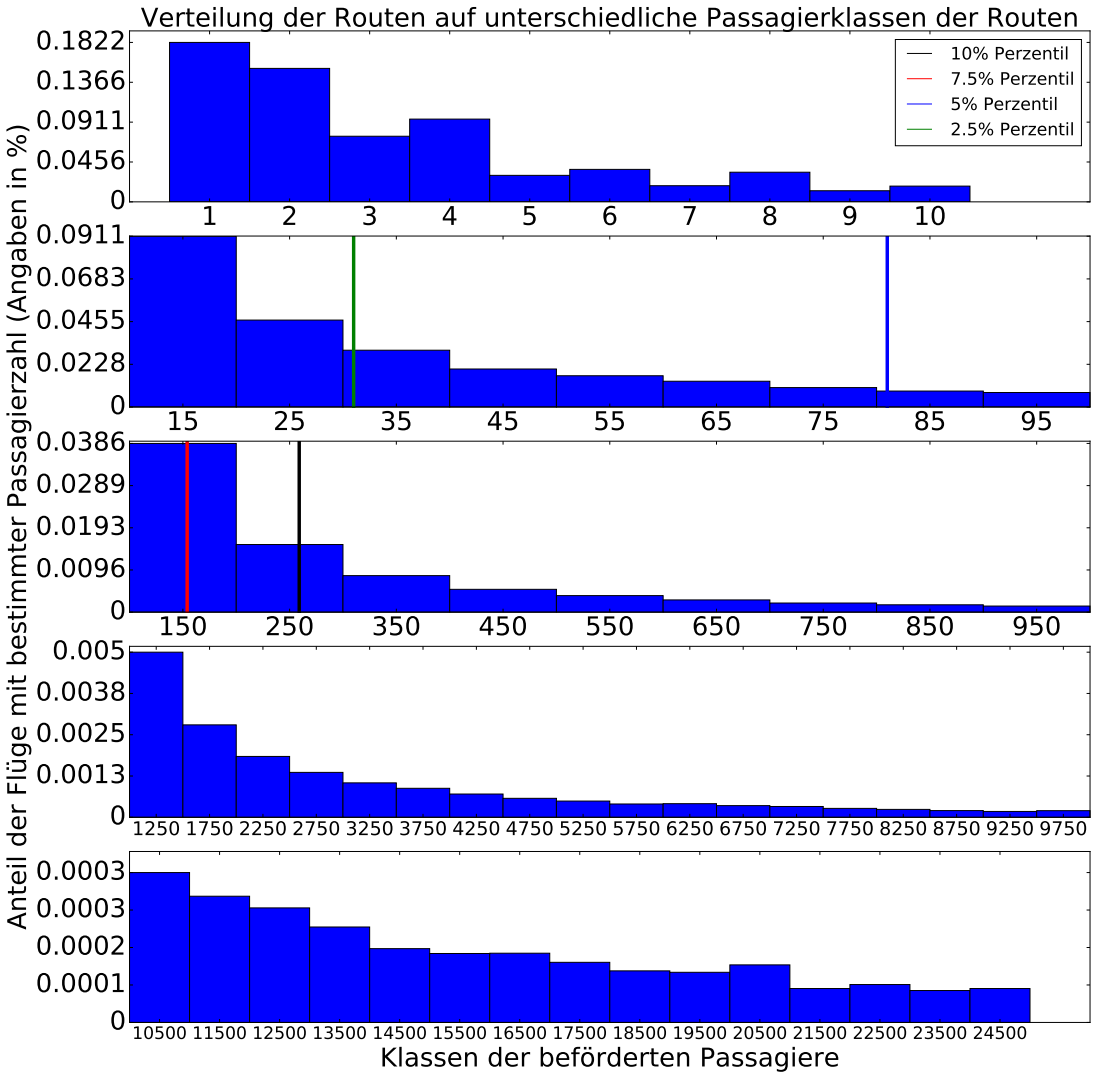


Abbildung 3.2: Histogramm der prozentualen Anzahl der Routen unterschiedlicher Passagierklassen. Die Zahl unter jedem Balken repräsentiert ein Intervall von Passagierzahlen, die zu einer Klasse zusammengefasst werden. Der Prozentsatz ergibt sich aus dem Quotienten der Anzahl der Routen, die in eine bestimmte Passagierklasse fallen und der Gesamtanzahl der Routen

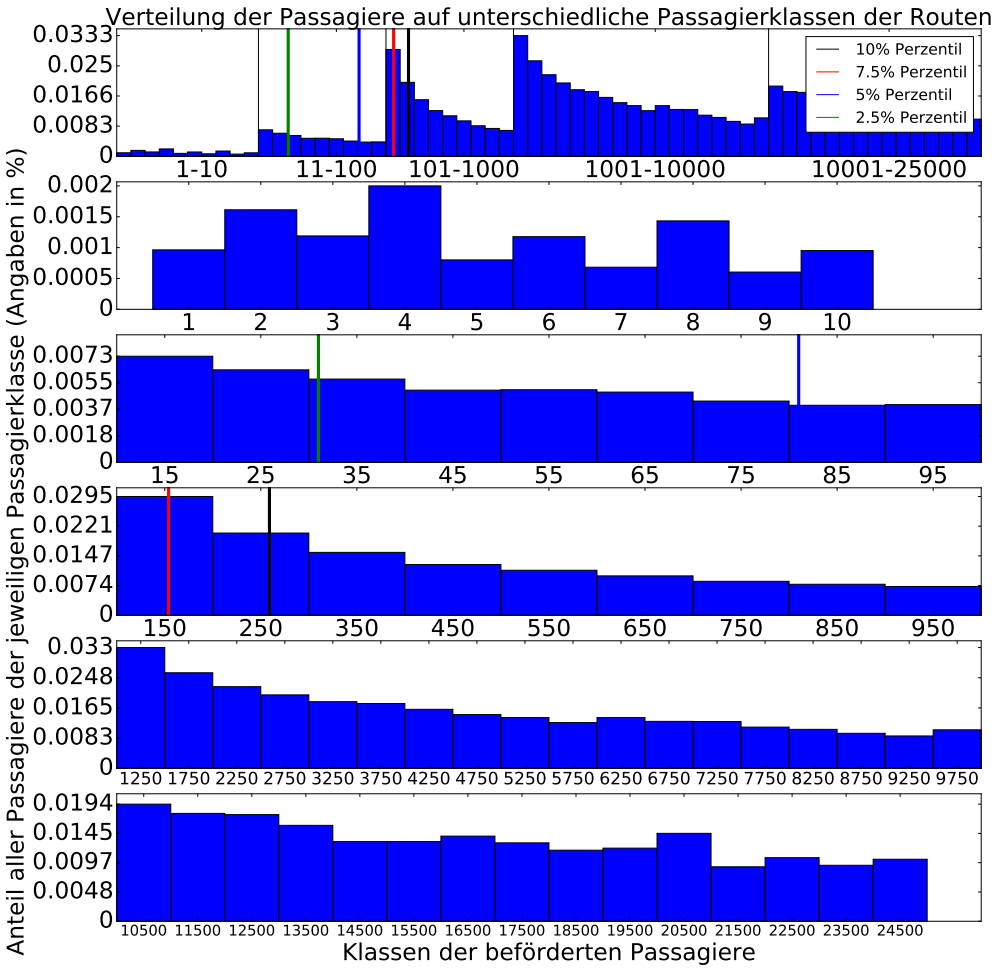
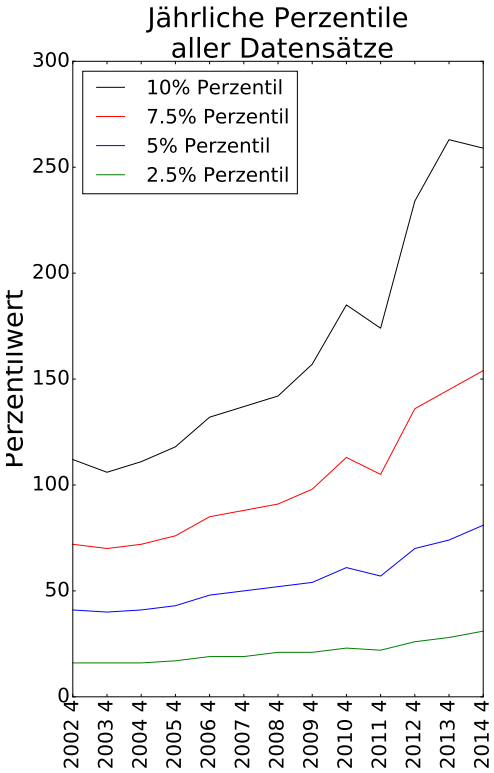
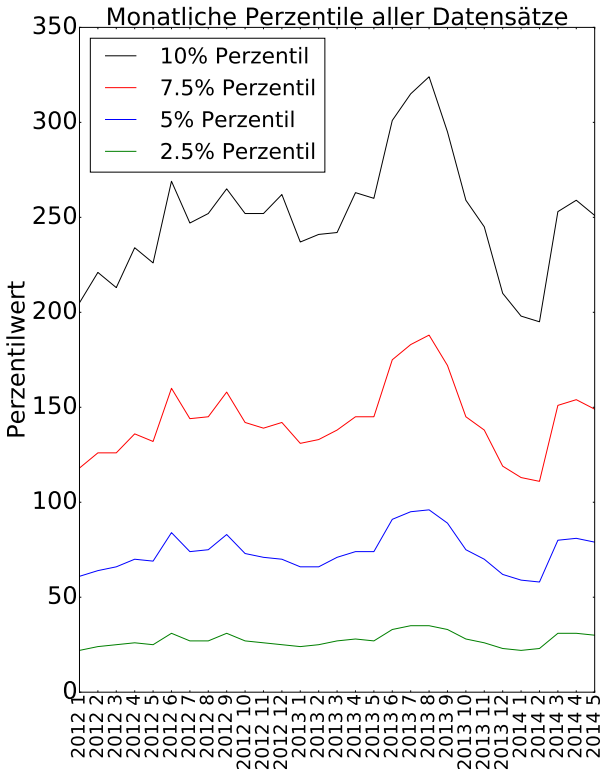


Abbildung 3.3: Histogramm der prozentualen Anzahl der Passagiere unterschiedlicher Passagierklassen. Die Zahl unter jedem Balken repräsentiert ein Intervall von Passagierzahlen, die zu einer Klasse zusammengefasst werden. Der Prozentsatz ergibt sich aus dem Quotienten der Gesamtpassagiere aller Routen, die in eine bestimmte Passagierklasse fallen und der Gesamtanzahl aller Passagiere. Der erste Abschnitt ist eine Zusammenstellung der unteren Abschnitte in einem Bild, wobei zu beachten ist, dass die Intervallbreite der Klassen beibehalten worden ist.

Die beiden nachfolgenden Abbildungen 3.4 und 3.5 zeigen die Wirkung der Perzentile auf die behaltenen Daten. In Monaten mit hohen Perzentilwerten sinkt die Anzahl der nicht entfernten Routen und umgekehrt, was deutlich in den Jahren bis 2012 und gegen Anfang 2014 zu sehen ist. Die Wahl des 5%-Perzentil bestätigt sich erneut, da die Ausprägungsstärke der Schwankungen der Perzentilrealisierungen und der Anzahl der behaltenen Routen bei 5% eher zur Übereinstimmung neigen, als bei 7.5% oder 10%. 2.5% wurde des Vergleichs wegen mit aufgeführt.



(a) Perzentile der jährlichen Daten für den Monat April

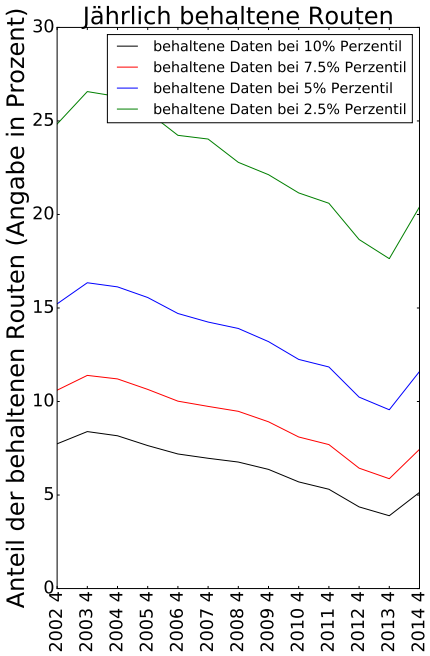


(b) Perzentile der monatlichen Daten ab Anfang 2012 bis Mai 2015

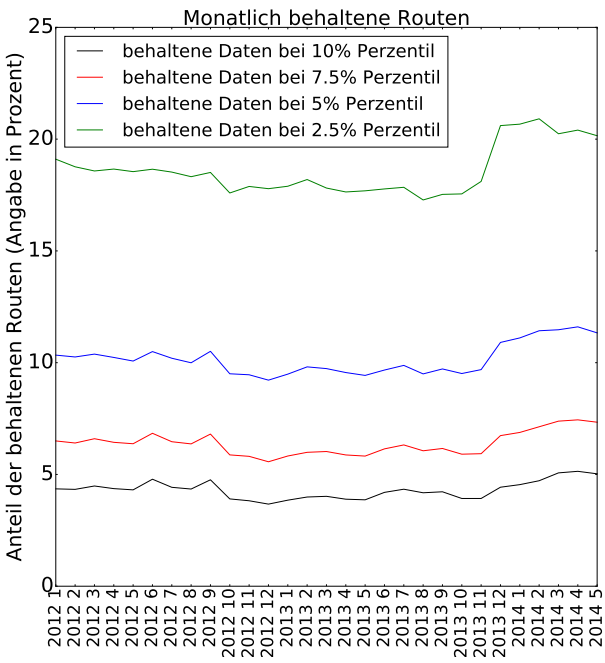
Abbildung 3.4: Perzentile für unterschiedliche Betrachtungszeiträume

Saisonbereinigung vs. jährliche Daten

Bei der Verwendung monatlicher Daten ist mit dem Problem zu rechnen, dass es Routen gibt, die zum Beispiel im Sommer hochfrequentiert sind und im Winter kaum einen Passagier aufweisen können. Dieses Saisonproblem tritt im nördlichen Raum der Erde häufig während der Sommerzeit und global vor allem während der Schulferien auf. Im asiatischen Raum ist in der Zeit des zwischen dem 21. Januar und dem 21. Februar stattfindenden chinesischen Neujahrsfestes mit erhöhtem Reiseaufkommen zu rechnen. Es ist ein globales Großreiseereignis mit China als Ziel für die Hinreise und als Startpunkt für die Rückreise. Würden solche hochsaisonalen Daten also als Referenz für einen Monat der Nebensaison verwendet, so ist mit Abweichungen zu rechnen. Zum Beispiel könnten Routen auftreten, die nur während der Hochsaison geflogen werden und dementsprechend viele Passagiere würden an diese Route gebunden sein. Andererseits können auch Routen nicht auftreten, die nur in der Nebensaison geflogen werden, aber aufgrund ihrer vergleichsweise geringen Frequentierung aus dem Vorhersageprozess entfernt wurden.



(a) Prozentsatz der behaltene Routen der jährlichen Daten für den Monat April



(b) Prozentsatz der behaltene Routen der monatlichen Daten ab Anfang 2012 bis Mai 2015

Abbildung 3.5: Prozentsatz der behaltene Routen für unterschiedliche Betrachtungszeiträume

Um derartige gleichmäßige Schwankungen in einer Art durchschnittlicher Größe betrachten zu können, ist die Anwendung einer sogenannten Saisonbereinigung denkbar. Sie ist eine statistische Methode aus dem Gebiet der Zeitreihenanalyse. Zur Information sei [SBA1], [SPE04], [SBA2] und [SBA15] empfohlen.
Die Erstellung einer Saisonbereinigung würde allerdings den Rahmen dieser Arbeit überschreiten. Im Einvernehmen mit den Betreuern wurde ein anderer Weg gewählt, der das oben beschriebene Problem umgeht. Dies geschieht mit der Verwendung jährlicher Daten. Sollte also eine Vorhersage über den Februar des Jahres 2016 verlangt sein, so werden zum Training des Modells lediglich die Februarmonate der vorangegangenen Jahre zur Verarbeitung herangezogen.
Die Saisonbereinigung lässt Daten zur Berechnung zu, die zeitlich nah am vorherzusagenden Ereignis liegen. Der Nachteil der jährlichen Daten besteht in ihrem Alter. So können Routen zeitlich weit zurückliegender Daten schon längst nicht mehr existent sein. Ebenso ist die Eröffnung neuer Routen oder gar Flughäfen denkbar, womit sich der Passagierfluss verlagert haben kann.
Sollte ein Benutzer der Modelle dennoch monatliche Daten gebrauchen wollen, so sei ihm angeraten, das BV4.1 Verfahren des Statistischen Bundesamtes der Bundesrepublik Deutschland in Form eines laut eigener Aussage einfach zu handhabenden Programmes zu benutzen [SBA3].

3.2.2 Kombination der dynamischen und statischen Daten

In Abschnitt 3.1 wurden verschiedene Dateien vorgestellt, die Datensätze bezüglich der Flugzeuge, Flughäfen, Routen und Wege beinhalten. Dieser Abschnitt stellt die Ermittlung der Inputinformationen jeder einzelnen Route für die Modelle vor.
Insgesamt werden 23 Parameter erzeugt. Diese teilen sich auf in 15 **metrische Variablen** (sie beschreiben konkrete Zahlenwerte) und 8 **kategorielle Variablen** (es werden Zahlen verwendet, um Kategorien darzustellen). Die kategoriellen Variablen sind binär codiert, wobei 1 für Ja und 0 für Nein steht.
Die Parameter sind für die Routen jedes Monats aufstellbar, der DD und DPD aufweisen kann. Wenn nicht, stehen sie nicht für die Lösung des Primär- oder Sekundärproblems zur Verfügung. Sollten einmal keine Informationen vorliegen, so wird ein entsprechender Eintrag „no entry“ eingefügt. Denn nicht alle Modelle benötigen alle Informationen und die reine Existenz des Eintrags ist bereits eine Information. Sollten alle Parameter benötigt werden, sind diejenigen Einträge nicht zu verwenden, die „no entry“ beinhalten.
In der folgenden Auflistung (Tabelle 3.25) werden zuerst die Variablennamen, eine kurze Beschreibung des Wertes und eine Auflistung der möglichen Werte angegeben.

Nr.	Variable	Beschreibung	Min. Wert	Max. Wert
		Metrische Parameter		
1	x_{a1}	Anzahl der Passagiere auf der Route	1	150 000
2	x_{a2}	Relative Wahrscheinlichkeit der Passagiere auf Route $= \frac{\text{Passagierzahl auf Route}}{\text{Passagierzahl auf Weg}}$	0.0	1.0
3	x_{a3}	Anzahl der Passagiere des zur Route gehörigen Weges	1.0	500 000
4	x_b	Durchschnittlich benötigte Zeit für diese Route in Stunden $= \sum_{\text{Segment}\in\text{Route}}$ durchschnittliche Segmentzeit	0.0	40.0
5	x_c	Frequenz der Route (wie oft wird sie geflogen) (Minimum über den einzelnen Segmentsummen)	1	1 000
6	x_{d1}	Maximale Hubgröße zwischen Origin und Destination nach FAA-Richtlinien (0, wenn Direktflug, die Wichtigkeit der Information geht auf die Direktflugvariable über)	1	5
7	x_{d2}	Maximale Hubgröße zwischen Origin und Destination nach Berechnungsmodell 1 der FAA	0.0	0.07
8	x_{d3}	Maximale Hubgröße zwischen Origin und Destination nach Berechnungsmodell 2 der FAA	0.0	0.03
9	x_e	$\frac{\text{Routenlänge}}{\text{kürzeste Routenlänge des Weges}}$	1.0	110.0
10	x_f	Anteil der Strecke, die bei Interkontinentalflug auf einem Kontinent geflogen wird (Angabe in Prozent)	0.0	1.0
11	x_g	Anzahl der Umstiege nach dem längstem Segment	0	3
12	x_h	Anzahl der Zwischenlandungen	0	3
13	x_i	Anzahl der Länder	1	5
14	x_j	Anzahl der Bahnhöfe zwischen Start und Ziel	0	3
15	x_k	Anzahl alternativer Routen	0	75
		Kategorielle Parameter		
16	x_{l1}	Minimale Flugzeugtyp, der auf den Segmenten der Route eingesetzt wird	0	15
17	x_{l2}	Maximale Flugzeugtyp, der auf den Segmenten der Route eingesetzt wird	0	15
18	x_{aa}	Ist es ein Direktflug?	0: Nein	1: Ja
19	x_{bb}	Ist es ein Internationaler Flug?	0: Nein	1: Ja
20	x_{cc}	Ist es ein Interkontinentalflug?	0: Nein	1: Ja
21	x_{dd}	Ist **kein** Flughafen mit kleiner Hubzahl zwischen Start- und Zielflughafen?	0: Nein	1: Ja
22	x_{ee}	Wenn internationaler Flug: mindestens 2x im Startland (0, wenn es kein internationaler Flug ist)	0: Nein	1: Ja
23	x_{ff}	Wenn internationale Flug: mindestens 2x im Zielland (0, wenn es kein internationaler Flug ist)	0: Nein	1: Ja

Tabelle 3.25: Auflistung aller verwendbaren Variablen mit kurzer Beschreibung und Angabe des Werteintervalls durch Minimal- und Maximalwert

Bemerkung 3.2.1
Die Parameter x_{a1}, x_{a2} und x_{a3} gehen nicht in die Berechnung ein, sondern bilden einzeln oder gemeinsam den Output y aus Abschnitt 3.1.1.

Erläuterungen zu den Variablen

Die Bedeutung der meisten Parameter ist offensichtlich. Bei einigen sind jedoch nähere Erläuterungen von Nöten. Die Berechnungsformeln entstammen [CON04], [DOY05], [GUI05], [FAA08] und [BUR03]:

6 : x_{d1}) Der Hubstatus x_{d1} eines Flughafens ist ein Maß für seine Wichtigkeit im internationalen Flughafenvergleich. Auf verschiedene Arten wird der Durchfluss der Passagiere gemessen. Bei diesem Parameter wurde die Standardeinteilung der Federal Aviation Administration (FAA) der Bundesluftfahrtbehörde der Vereinigten Staaten angewendet. Die FAA teilt Flughäfen in sechs Kategorien (Large, Medium, Small, Nonhub Primary, Nonhub Nonprimary, Too small) bezüglich des Passagierstroms auf. Dabei wird für jeden Flughafen i gezählt, wie viele Menschen ihn als Start-, Ziel- oder Umstiegsflughafen benutzt haben. Die Zahl sei in diesem Kontext einfach mit x_i bezeichnet. Die Passagiere aller Routen werden summiert. Flughäfen, auf die wenigstens 1% dieser Gesamtpassagierzahl entfällt, erhalten das Prädikat „Large“. Sei I die Menge aller Flughäfen.

„Large“:	$x_i \geq 1\%$
„Medium“:	$1\% > x_i \geq 0.25\%$
„Small“:	$0.25\% > x_i \geq 0.05\%$
„Nonhub Primary“:	$0.05\% > x_i \geq 10\,000$
„Nonhub Nonprimary“:	$10\,000 > x_i \geq 2\,500$
„Too small“:	$2\,500 > x_i$

Die konkreten Zahlwerte sind konkrete Passagierzahlen.
Da davon auszugehen ist, dass Start- und Zielflughafen lediglich Zubringer sind, wird der maximale Hubstatus der dazwischenliegenden Flughäfen bestimmt. Sollte es sich um einen Direktflug handeln, so wird der Wert auf 0 gesetzt, die Wichtigkeit der Information geht auf den Direktflugparameter über. Abbildung 3.6 zeigt die den prozentuale Verteilung der Flughäfen auf die verschiedenen Hubstadien.

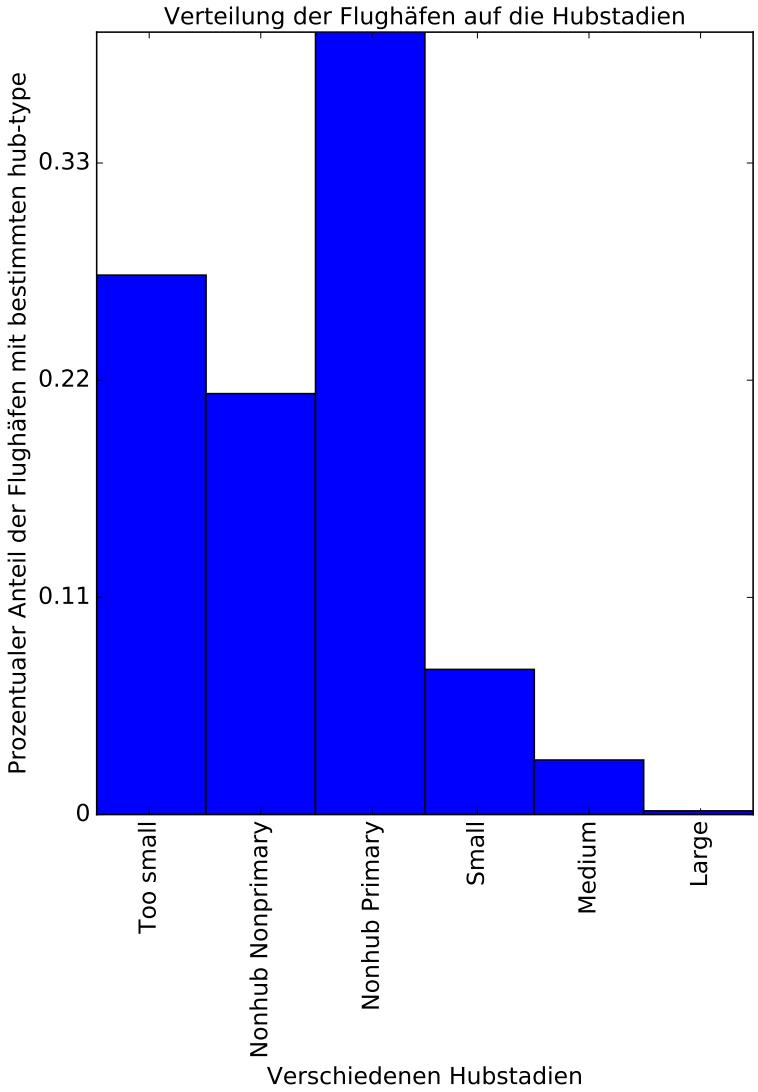


Abbildung 3.6: Verteilung der Flughäfen auf die Hubstadien nach x_{d1}

7 : x_{d2} und 8:x_{d3}) Auch hier wird der Hubstatus von Formeln bestimmt, die bei der FAA Anwendung finden. Im Gegensatz zu x_{d1} gibt es jedoch keine Klasseneinteilung, sondern es werden konkrete reelle Werte berechnet. Weiterhin wird zwischen Passagieren unterschieden deren Start oder Ziel der Flughafen i ist (gezählt mit Variable od_i) und Passagieren, die Flughafen i lediglich als Zwischenstop benutzen (gezählt mit Variable c_i). $P = \sum_{i \in I} od_i$ beschreibe die Gesamtzahl der Flugpassagiere. Dann sind folgende drei Kennzahlen bestimmbar

- $OD_i = \frac{od_i}{P}$ Indikator für Wichtigkeit des Flughafens i als Verkehrserzeuger
- $C_i = \frac{c_i}{P - od_i}$ Flussbasierter Indikator für Wichtigkeit des Flughafens als Verbindungspunkt
- $E = \sum_{i \in I} (od_i + c_i)$ als totale Anzahl der Reisenden im Flugnetzwerk.

Dann ergibt sich der einfache Hubstatus von Flughafen i als

$$x_{d2} = \frac{od_i + c_i}{E}, \tag{3.3}$$

also als Anteil der Passagiere, die i genutzt haben, im Vergleich zu den Passagieren, die überhaupt Flughäfen genutzt haben (E).
Eine zusammengefasste Aufstellung von Passagieren, die i als Start oder Ziel, und Passagieren, die i lediglich zum Umsteigen nutzen, liefert

$$x_{d3} = OD_i \frac{P}{E} + C_i \frac{P - OD_i}{E}. \tag{3.4}$$

Zur Vermeidung von Komplikationen sei darauf hingewiesen, dass alle genannten Variablenbezeichnungen nur für diesen Abschnitt gelten.
Es existieren weitere Kennzahlen für die Bedeutung eines Flughafens wie den früher verwendeten Gini-Index [REY98] und [GUI05] oder Zentralitätsindize aus dem Bereich der Netzwerke. Alternativ kann auch die Reisezeit als Index herangezogen werden [RED11].

10 : x_f) Den Großteil der Zeit bei einem Interkontinentalflug sollte der Flug von einem Kontinent zum anderen benötigen. In den meisten Fällen macht dies den Hauptteil der Reise aus. Zusätzlich lange Zubringerflüge zu haben, bedeutet größeren Stress für den Passagier. Damit folgt: Je geringer dieser Anteil, desto besser. Weiterhin gilt: Je höher die Anzahl der Zubringerflüge, desto größer das Risiko, dass der Interkontinentalflug nicht erreicht wird.

11 : x_g) Je mehr Anschlussflüge nach einem Zubringerflug, desto anstrengender ist die Reise. Kleine Zahlen sind an dieser Stelle also positiv einzuschätzen.

21 : x_{dd}) Der untersuchte Wert bezieht sich auf x_{d2}, der im Allgemeinen gröber als x_{d3} und feiner als x_{d1} ist. Sollte zwischen Start und Ziel ein Flughafen angeflogen werden, der einen kleineren Hubstatus besitzt als der Start- oder Zielflughafen, dann ist davon auszugehen, dass die Route insgesamt unwichtig ist. Grund dafür ist die Annahme, dass es sich bei wenigstens einem der beiden Start- und Zielflughäfen um einen Zubringerflughafen handelt. Befindet sich auf der Route also ein noch unwichtigerer Flughafen, so wird die Route aller Voraussicht nach nicht oft benutzt.

22 : x_{ee}) Ein Umstieg vor dem Verlassen des Startlandes kann zeitliche Vorteile bezüglich des Gepäcks und der Einreisekontrolle haben. Die Kontrollen bei einem Inlandsflug sind geringer, das Aufgeben des Gepäcks ist weniger umständlich. Wird dann der Flughafen erreicht, der den Passagier in das nächste Land bringt, hat er keine Kontrollen und Umständlichkeiten des Gepäcks wegen zu befürchten. Natürlich ist dieser Umstand stark von den jeweiligen gesetzlichen Rahmenbedingungen der betroffenen Länder abhängig.

23 : x_{ff}) Argumentation analog zu x_{ee}

4 Allgemeine Vorbetrachtungen

Dieses Kapitel liefert einen groben Überblick über die Probleme und auftretenden Fehler die bei der Wahl eines geeigneten Modells auftreten. Es bedient sich dabei noch nicht der im vorherigen Kapitel 3 konkret beschriebenen Daten, sondern hält sich allgemein. Der Einsatz der Daten erfolgt erst ab Kapitel 5.
Dieses Kapitel orientiert sich dabei teilweise an [HAS08] sowie an [KAL], [MUE], [KOM], [BVD] und [SRM]. Die aufgeführten Abbildungen entstammen diesen Quellen, vor allem diejenigen aus Abschnitt 4.4.3 sind den Originalen fast 1:1 nachempfunden.

4.1 Input, Output

Nach der in Abschnitt 2.2 vorgestellten Aufgabenstellung sind Verfahren zu erstellen, die auf Prognosemethoden basieren. Die Prognosemodelle nehmen Daten entgegen, verarbeiten sie und liefern dem Anwender ein Ergebnis.
Die eingegebenen Daten heißen **Input** („Eingabe“, „Features“, „unabhängige Variablen“ oder auch „Predictors“) und werden allgemein mit der Variable $\mathbf{X}$ symbolisiert. Im Falle der ausgegebenen Daten beziehungsweise des Ergebnisses lautet die hier verwendete Bezeichnung **Output** („Ausgabe“). Andere in der Literatur übliche Varianten sind „Responses“ oder „abhängige Variablen“. Ihre Symbolisierung erfolgt durch die Variable $\mathbf{Y}$.
Wie in Abschnitt 3.1 aufgeführt ist, bilden die DD den Input X und die DPD den Output Y. Ein Paar (X, Y) zusammengehöriger Input- und Outputdaten wird **Beobachtung** genannt. Aufgrund der Existenz von Y können Prognoseverfahren des überwachten Lernens verwendet werden. Mehr dazu findet sich in Abschnitt 4.2.
Bei den Daten erfolgt die Unterscheidung in zwei wesentliche Kategorien: Den **Quantitativen Daten** und den **Qualitativen Daten**. Bei den vorliegenden Daten handelt es sich um quantitative Daten. Sie sind numerischer Natur. Üblicherweise ist eine euklidische Metrik auf ihnen definiert und sie unterliegen einer Ordnungsrelation, sodass nah beieinanderliegende Inputdaten ähnliche Outputdaten erzeugen. Qualitative Daten dagegen sind nicht numerisch. Meist handelt es sich um Kategorien, sodass keine Metrik auf ihnen definiert ist. Beispiele dafür sind Farbeinteilungen, krank oder nicht krank, etc.. Manche sind numerisch kodierbar. Bei anderen wiederum lässt sich eine gewisse Ordnungsrelation finden (klein, mittel, groß). Prognoseverfahren des überwachten Lernens für quantitative Daten nennen sich **Regression**. Bei qualitativen Daten ist von **Klassifikation** die Rede.

4.2 Training und Überwachtes Lernen

Die ab Kapitel 5 vorgestellten Regressionsverfahren bestehen aus einer Schätzfunktion $\pi : X \to [0,1]$, welche einen geschätzten Output $\widehat{Y} = \pi(X)$ ermittelt. Weiterhin wird von einem parametrischen Schätzmodell $f_\theta(x) = E[Y|X = x]$ ausgegangen. Von diesem aus erfolgt der Übergang zu $Y = f(X) + \varepsilon$ (siehe nächsten Abschnitt 4.3) und der Schätzung $\pi(x) = E[Y|X = x]$. Die Regressionsverfahren unterscheiden sich in der Wahl der Schätzfunktion π.
Im Allgemeinen beinhaltet π einen oder mehrere Funktionsparameter oder auch Anpassungsparameter, welche in einem Parametervektor, hier $\boldsymbol{\beta}$ genannt, zusammengefasst werden. In der Fachliteratur wird statt $\boldsymbol{\beta}$ oft θ verwendet. Ein Nutzer der Regression hat neben der Wahl einer günstigen Regressionsfunktion das Problem der optimalen Einstellung der Funktionsparameter. Neben der Anwendung von π ist daher oft ein vorangehendes Verfahren nötig, welches diese Einstellung vornimmt (daher auch der Zweitname „Anpassungsparameter“). Der Prozess der Anpassung heißt **Training** oder auch „Lernen“. Unter Verwendung älterer, bereits bekannter Daten sind dabei die Werte von $\boldsymbol{\beta}$ derart zu ermitteln, dass die Differenz von Y und $\widehat{Y}$ oder ein bestimmtes Fehlermaß minimal wird. Genaueres dazu ist in Abschnitt 4.4.4 beschrieben. Wird von einem Regressionsverfahren gesprochen, so ist meist das Training und die Anwendung gemeint.

Training lässt sich in zwei Hauptmethoden unterscheiden: **Lernen mit** und **Lernen ohne Lehrer**, beziehungsweise **Überwachtes Lernen** und **Unüberwachtes Lernen** („Supervised/Unsupervised Learning“). Überwachtes Lernen anwendbar, wenn zu einem anzupassenden Regressionsmodell sowohl die Input- als auch die gewünschten Outputinformationen der Trainingsdaten bekannt sind. Sollten nur die Inputinformationen vorliegen, muss das Modell mittels Unüberwachtem Lernen angepasst werden. Bei dem behandelten Problem ist Überwachtes Lernen anwendbar, da sowohl die DD als Input-, als auch die DPD als Outputinformationen bis zurück ins Jahr 2002 vorliegen. Daher ist es nicht nötig, das Unüberwachte Lernen zu beschreiben. Es sei hierbei aber auf [UNS1] und [UNS2] verwiesen.
Die Anpassung oder auch das Training mittels Überwachtem Lernen erfolgt nun, indem das anzupassende Modell mit Startwerten von $\boldsymbol{\beta}$ initialisiert wird. Anschließend werden die Inputdaten eingegeben und der geschätzte Output $\widehat{Y}$ erfasst. Da der tatsächliche Output Y bekannt ist, können mittels einer Fehlerfunktion und weiterführender Verfahren Rückschlüsse gezogen werden, wie die Schätzparameter $\boldsymbol{\beta}$ zu verändern sind, um den Fehler der Fehlerfunktion zu minimieren. Die Funktion ist in der Praxis oft die des **Mittleren Quadrierten Fehlers**.
Wie die Trainingsdaten zu wählen sind, ob und wie oft dieser Vorgang wiederholt werden sollte ist in 4.4.4 beschrieben.

Neben der Wahl des richtigen Anpassungsverfahrens ist die Wahl der Fehlerfunktion eine der wichtigsten theoretischen Grundvoraussetzungen für das Training. Eine andere

Fehlerfunktion wäre der Mittlere Betragsmäßige Fehler. Aufgrund ihrer Unstetigkeitsstelle wird sie aber oft nicht verwendet. Der bereits erwähnte Mittlere Quadrierte Fehler besitzt dieses Problem nicht, ist differenzierbar (wenn auch π dies ist) und neben eben genannter Betragsfunktion die einfachste mögliche Fehlerdarstellung. Sie bildet daher die Grundlage für die meisten Regressionsverfahren, weshalb sie auch in dieser Arbeit verwendet wird. Zu weiteren Vor- und Nachteilen, sowie einer Aufstellung weiterer gebräuchlicher Fehlerfunktionen sei auf [AGR02] verwiesen.

4.3 Additives Fehlermodell

Für jedes Modell und alle Daten existieren Einflüsse unbekannter Natur, die einen gewissen unbeeinflussbaren und unbestimmbaren Grundfehler ε erzeugen. Ein solcher Fehler wird in den Naturwissenschaften als **Weißes Rauschen** bezeichnet und findet oft in der Statistik und Prognose Anwendung, um Störungen in einem ansonsten idealen Modell darzustellen. Es ist als diskreter, schwach stationärer, stochastischer Prozess unkorrelierter Zufallsvariablen zu beschreiben. Sie besitzen einen Erwartungswert $\mathrm{E}[\varepsilon] = 0$ und eine konstante Varianz.
Die einfachste Art und Weise, das weiße Rauschen in die Regressionsverfahren zu integrieren ist die Annahme des **Additiven Fehlermodells**

$$\varepsilon(w) := -f(X(w)) + Y(w) \tag{4.1}$$

mit $f(x) = \mathrm{E}[Y|X = x]$, wobei die bedingte Verteilung $P(Y|X)$ von X nur durch den bedingten Mittelwert $f(x)$ abhängt. Die Annahmen des Modells beziehen sich auf die Annahmen für ε.
Das additive Fehlermodell ist exakt. Den approximativen Charakter bilden die Annahmen an ε. Es ist eine nützliche Darstellung des unbekannten wahren Modells, da die meisten Beobachtungen keine funktionale Beziehung $Y = f(X)$ haben werden. Im Gegenteil, es existieren, wie eingangs genannt, oft weitere unmessbare Einflussvariablen. Unter anderem auch der Messfehler. Das additive Fehlermodell vereinigt all diese Abweichungen von einer deterministischen Input-Output-Beziehung im Fehler ε.
Die Annahme der unabhängig und identisch verteilten Fehler in Gleichung (4.1) ist nicht streng notwendig, sollte aber auch nicht aus den Augen verloren werden. Das Modell verträgt sich mit der Verwendung des mittleren quadrierten Fehlers.
Zur Vermeidung der Unabhängigkeitsannahme kann folgende Modifikation der Varianz vorgenommen werden, sodass sowohl der Mittelwert als auch die Varianz von X abhängen: $\mathrm{Var}(Y|X = x) = \sigma(x)$. Im Allgemeinen kann die bedingte Verteilung $P(Y|X)$ auf sehr komplizierte Art und Weise von X abhängen, was das additive Fehlermodell jedoch ausschließt.

4.4 Bias-Varianz-Dilemma

Nach der Einführung des weißen Rauschens, des Fehlers unmessbarer Einflüsse, stellt sich nun die Frage, welche weiteren messbaren Fehler existieren und wie sie eventuell vermieden werden können.
Dazu seien die Trainingsdaten X und Y betrachtet. Eine spezielle Beobachtung sei mit (x_i, y_i) bezeichnet, beziehungsweise eine konkrete Realisierung mit x und y. Es sei angenommen, diese Daten unterliegen dem additiven Fehlermodell $y_i = f(x_i) + \varepsilon$ mit $\varepsilon \sim N(0, \sigma^2)$. Mittels eines Lernalgorithmus ist eine Schätzfunktion zu $\widehat{f}(x)$ zu bestimmen, welche den strukturellen Zusammenhang $y = f(x)$ so gut als möglich approximieren soll. Aufgrund des weißen Rauschens ist eine optimale Annäherung nicht realisierbar. Zur Messung der Abweichung wird der mittlere quadratische Fehler $\mathrm{E}\left[y - \widehat{f}(x)\right]^2$ eingesetzt, wobei dieser Fehler für alle Beobachtungen (x_i, y_i) innerhalb, als auch für alle weiteren Beobachtungen außerhalb der Trainingsdaten zu minimieren ist.

Im Folgenden wird der erwartete mittlere Fehler in drei einzelne Fehlerterme zerlegt. Dafür sind einige kleine Vorbetrachtungen von Nöten.

Es gilt

$$\mathrm{E}\left[X^2\right] = \mathrm{Var}[X] + \mathrm{E}[X]^2. \tag{4.2}$$

Da $f(x)$ deterministisch ist, gilt $\mathrm{E}[f(x)] = f(x)$. Eingangs des Kapitels wurde die Annahme $y = f(x) + \varepsilon$ und $\mathrm{E}[\varepsilon] = 0$ festgelegt (wobei ε nun eine Zufallsvariable ist), daher gilt

$$\mathrm{E}[Y|X = x] = \mathrm{E}[y] = \mathrm{E}[f(x) + \varepsilon] = \mathrm{E}[f(x)] = f(x). \tag{4.3}$$

Weiterhin ist mit oben getroffene Annahme $\mathrm{Var}[\varepsilon] = \sigma^2$ gültig, was zu folgender Aussage führt

$$\begin{aligned} \mathrm{Var}[y] &= \mathrm{E}\left[(y - \mathrm{E}[y])^2\right] = \mathrm{E}\left[(y - f(x))^2\right] = \mathrm{E}\left[(f(x) + \varepsilon - f(x))^2\right] \\ &= \mathrm{E}\left[\varepsilon^2\right] = \mathrm{Var}[\varepsilon] + \mathrm{E}[\varepsilon]^2 = \sigma^2. \end{aligned} \tag{4.4}$$

Aufgrund der im letzten Abschnitt 4.3 getroffenen Unabhängigkeitsannahme von ε und

$\widehat{f}(x)$ folgt

$$\begin{aligned}
\mathsf{E}\left[\left(y-\widehat{f}(x)^2\right)^2\right] &= \mathsf{E}\left[y^2+\widehat{f}(x)^2-2y\widehat{f}(x)\right] \\
&= \mathsf{E}\left[y^2\right]+\mathsf{E}\left[\widehat{f}(x)^2\right]-\mathsf{E}\left[2y\widehat{f}(x)\right] \\
&= \mathsf{Var}\left[y\right]+\mathsf{E}\left[y\right]^2+\mathsf{Var}\left[\widehat{f}(x)\right]+\mathsf{E}\left[\widehat{f}(x)\right]^2-2f(x)\mathsf{E}\left[\widehat{f}(x)\right] \\
&= \mathsf{Var}\left[y\right]+\mathsf{Var}\left[\widehat{f}(x)\right]+\left(f(x)-\mathsf{E}\left[\widehat{f}(x)\right]\right)^2 \\
&= \mathsf{Var}\left[y\right]+\mathsf{Var}\left[\widehat{f}(x)\right]+\mathsf{E}\left[f(x)-\widehat{f}(x)\right]^2. \qquad (4.5)
\end{aligned}$$

Mit

$$\mathsf{Bias}\left[\widehat{f}(x)\right] = \mathsf{E}\left[\widehat{f}(x)-f(x)\right] \qquad (4.6)$$

und

$$\mathsf{Var}\left[\widehat{f}(x)\right] = \mathsf{E}\left[\left(\widehat{f}(x)-\mathsf{E}\left[\widehat{f}(x)\right]\right)^2\right] \qquad (4.7)$$

ergibt sich die sogenannte **Bias-Varianz-Zerlegung**

$$\mathsf{E}\left[\left(y-\widehat{f}(x)\right)^2\right] = \mathsf{Bias}\left[\widehat{f}(x)\right]^2+\mathsf{Var}\left[\widehat{f}(x)\right]+\sigma^2. \qquad (4.8)$$

Jeder dieser drei Terme steht für unterschiedliche Fehleraspekte des Schätzmodells:

- σ^2 vertritt den irreduziblen Fehler, welcher über das weiße Rauschen in das Modell einfließt. Dieser Fehler ist stets vorhanden und lässt sich nicht vermeiden. Er ist vergleichbar mit einer Art Grundlast oder Fixkosten, um Termini aus dem Bereich der Finanzmathematik zu verwenden. Er bildet die untere Fehlerschranke für ungesehene Testdaten. Umfasst σ^2 einen großen Anteil am mittleren quadratischen Fehler, so lässt sich auf die Existenz vieler unbekannter Einflüsse schließen. Es sollte überlegt werden, neue Daten mit mehr vertretenen Einflüssen zu organisieren. Dieser Term ist außerhalb der Kontrolle des Anwenders.
- Die **Varianz** beschreibt die Schwankungen der geschätzten Realisierungen $\widehat{f}(x)$ um den Erwartungswert. Bei Betrachtung der Trainingsdaten beschreibt die Größe der Varianz, wie gut sich das Modell an diese Daten angepasst hat. Danach sollte das Modell auf neue Daten, sogenannte Testdaten, angewendet werden, um zu evaluieren, wie es mit anderen Daten zurechtkommt, auf die es nicht angepasst wurde. Eine hohe Varianz im ersten Fall ist prinzipiell schlecht, denn sie besagt, das sich das Modell (noch) nicht ausreichend auf die Trainingsdaten angepasst wurde. Niedrige Werte im ersten und hohe Werte im zweiten Fall lassen auf eine sogenannte **Überanpassung** („Overfitting“) schließen. Dies bedeutet, das sich das Modell nahezu exakt an die Trainingsdaten angepasst hat und nicht in der Lage ist, den richtigen Output zu von den Trainingsdaten abweichenden In-

putwerten zu erzeugen. Die Trainingsdaten wurden faktisch „auswendig gelernt“. Eine Modellanpassung mit sehr guter Varianz im ersten und zweiten Fall ist zwar nicht auszuschließen, in der Praxis jedoch so gut wie nicht anzutreffen. Es muss ein Kompromiss aus nicht optimaler Varianz bei den Trainingsdaten und dafür verbesserter Varianz bei den Testdaten gefunden werden.

- Der **Bias** oder auch „systematischer Fehler/Verzerrung“ bezieht sich auf das verwendete Modell selbst und beschreibt, wie wenig erwartungstreu die Schätzfunktion ist, also wie stark sie im Mittel vom zu schätzenden Wert abweicht. Vereinfacht gesprochen liefert der Bias eine Art Maß dafür, wie stark die Struktur des Schätzmodells $\widehat{f}$ von der wahren Natur des Modells f abweicht, also wie genau das wahre Modell getroffen wurde. Je besser die Annahmen sind, die über das Modell getroffen werden, desto besser, also niedriger, der Bias. Wird beispielsweise bei einer eindimensionalen Funktion bei wachsendem x von einer wachsenden Funktion ausgegangen, obwohl der Verlauf in Wirklichkeit fallend ist, so ist der Biaswert sehr hoch. Sollte die Input-Output-Beziehung allerdings exakt getroffen sein, so ist der Bias gleich 0.

Nach der Beschreibung der drei Terme soll nun das eigentliche Dilemma beschrieben werden. Das Ziel der Wahl und des Trainings eines Schätzmodells ist die Auffindung einer möglichst korrekten Abbildung der Beziehung der Beobachtungen der Trainingsdaten bei gleichzeitig guter Generalisierung für die unbekannten Testdaten. Der irreduzible Fehler sei vernachlässigt, da er nicht geändert werden kann. Gesucht wird also ein Modell mit geringem Bias und geringer Varianz. Es ist üblicherweise nicht möglich, ein Schätzverfahren zu finden, welches beide Bedingungen erfüllt. Meist ist der Anwender zu einer Wahl aus Modellen mit hohem Bias und niedriger Varianz oder niedrigem Bias und hoher Varianz gezwungen.

Modelle mit hohem Bias und niedriger Varianz zeichnen sich oft dadurch aus, dass sie relativ einfach und deshalb auch für schwierigere Daten anwendbar sind (obwohl dabei auch Probleme auftreten können, wie in den nächsten Abschnitten zu sehen ist). Die Beziehungen in den Daten wird dabei aber oft nicht gut abgebildet, sodass diese Modelle wichtige Gesetzmäßigkeiten außer Acht lassen und zur Unteranpassung neigen können.
Demgegenüber stehen die Modelle mit niedrigem Bias und hoher Varianz. Sie zeichnen sich oft dadurch aus, dass sie komplex sind und die Beziehungen in den Trainingsdaten gut darstellen können. Sie treffen dabei viele und/oder starke Annahmen und Voraussetzungen bezüglich der Daten und/oder enthalten komplexe Funktionen. Leider wird dabei meist auch ein Gutteil des weißen Rauschens mit abgebildet, was in ungenauen Vorhersagen auf den Testdaten mündet. Ein Beispiel für ein solches Schätzverfahren ist das **k-nearest-neighbor**-Modell.
Die Frage nach einem günstigen Modell liegt also in der **Modellkomplexität**. Sie entspricht der Anzahl der Modellparameter, auch **Freiheitsgrade** genannt.

Hierbei gilt das **No-free-Lunch**-Theorem, welches besagt, dass kein Modell besser oder schlechter als ein anderes ist. Die Güte ist allein abhängig von der getroffenen Auswahl der Trainingsdaten, welches je nach Modell zu einem günstigen oder ungünstigen Ergebnis führen können.

Wie aus der obigen Beschreibung ersichtlich, liegt das Dilemma darin, ein Modell und einen Grad der Anpassung zu finden, dass die Summe aus Bias und Varianz ein Minimum erreicht.

Für eine anschaulichere Verdeutlichung des Problems seien die Schätzverfahren für die Extremszenarien aus Bias und Varianz vorgestellt: Das k-nearest-neighbor-Modell und das lineare Modell.

4.4.1 Lineares Modell

Das lineare Modell wird ausführlich in Abschnitt 5.2 behandelt, daher ist das vorgestellte Verfahren lediglich grob dargestellt.

Es bestehe jedes Inputpaar $(\boldsymbol{x}_i, y_i)$ der Trainingsmenge der Mächtigkeit N aus einem Inputvektor $\boldsymbol{x}$ mit p verschiedenen Eigenschaften und einem Outputwert y. Alle Daten werden in einer Inputmatrix $\boldsymbol{X}$ und einem Outputvektor $\boldsymbol{y}$ dargestellt. Die Bewertung der Daten erfolgt über den Parametervektor $\boldsymbol{\beta}$, welcher ein zusätzliches Element β_0 für das weiße Rauschen enthält. Durch Integration der Konstanten 1 in allen $\boldsymbol{x}$ kann das lineare Modell folgendermaßen dargestellt werden

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}. \tag{4.9}$$

Die Lösung erfolgt über die **Methode der kleinsten Quadrate**:
Dabei wird die Fehlerfunktion der mittleren quadratischen Abweichung

$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}\right)^2 \tag{4.10}$$

nach $\boldsymbol{\beta}$ abgeleitet, die erste Derivate gleich 0 gesetzt und nach $\boldsymbol{\beta}$ umgestellt. Das Ergebnis ist eine exakte Berechnungsvorschrift zur Schätzung von $\widehat{\boldsymbol{\beta}}$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \tag{4.11}$$

mit der Voraussetzung, dass $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ nicht singulär ist.

Neben der Methode der kleinsten Quadrate (Minimierung des mittleren quadratischen Fehlers) existieren viele weitere Lösungsansätze. Je nach Modell können Nebenbedingungen per Lagrangeterm in die zu optimierende Funktion integriert werden. Dabei ist die Entstehung ungünstiger Lösungen möglich. So zum Beispiel sind Modelle mit Nebenbedingungen denkbar, bei denen der Lagrangeparameter gegen $-\infty$ geschätzt werden kann, um ein Minimum zu erreichen. Was bei vorliegendem Problem jedoch nicht der Fall ist.
Trotzdem kann stattdessen die allgemeinere Form der **Maximum - Likelihood - Schätzung** mit der Likelihoodfunktion $L(\boldsymbol{\beta}) = \sum_{i=1}^{N} P_{\boldsymbol{\beta}}(y_i)$ verwendet werden. Einer der oben erwähnten weiteren Lösungsansätze. Sie geht davon aus, das die beste Wahl eines $\boldsymbol{\beta}$ diejenige ist, welche die Likelihoodfunktion maximiert. Eine genauere Betrachtung erfolgt in Abschnitt 5.4.6.

Bemerkung 4.4.1
Dies ist ein Beispiel für ein Modell mit hohem Bias und geringer Varianz, wie in Abschnitt 4.4.3 verdeutlicht wird. Es setzt voraus, dass die Daten über eine globale lineare Funktion wohlapproximiert sind.

4.4.2 k-nearest-neighbor-Modell

Das k-nearest-neighbor-Modell („knn“, k-nächste-Nachbarn) ist recht simpel. Jedem neuen unbekannten Datenpunkt x wird der Output zugeordnet, der sich aus dem Mittelwert der Outputs y_i der umliegenden k nächsten Datenpunkte x_i ergibt. Bei einer endlichen Menge an möglichen Outputs sind Schwellwerte zu verwenden.

Wenn $N_k(x)$ die Menge der k nächsten Nachbarn um einen Datenpunkt x ist, so ergibt sich folgendes knn-Modell

$$\widehat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i. \tag{4.12}$$

Dieses Modell setzt voraus, dass sich die Datenpunkte in einem metrischen Raum befinden. Für das nachfolgende Beispiel kann von einem metrischen Raum ausgegangen werden.

Bemerkung 4.4.2
Das 1-nearest-neighbor Modell ist ein Beispiel für ein Verfahren mit geringem Bias und hoher Varianz, wie in Abschnitt 4.4.3 verdeutlicht wird. Es setzt voraus, dass die Daten über eine lokal konstante Funktion wohlapproximiert sind.

4.4.3 Vergleich des linearen und des k-nearest-neighbor-Modells

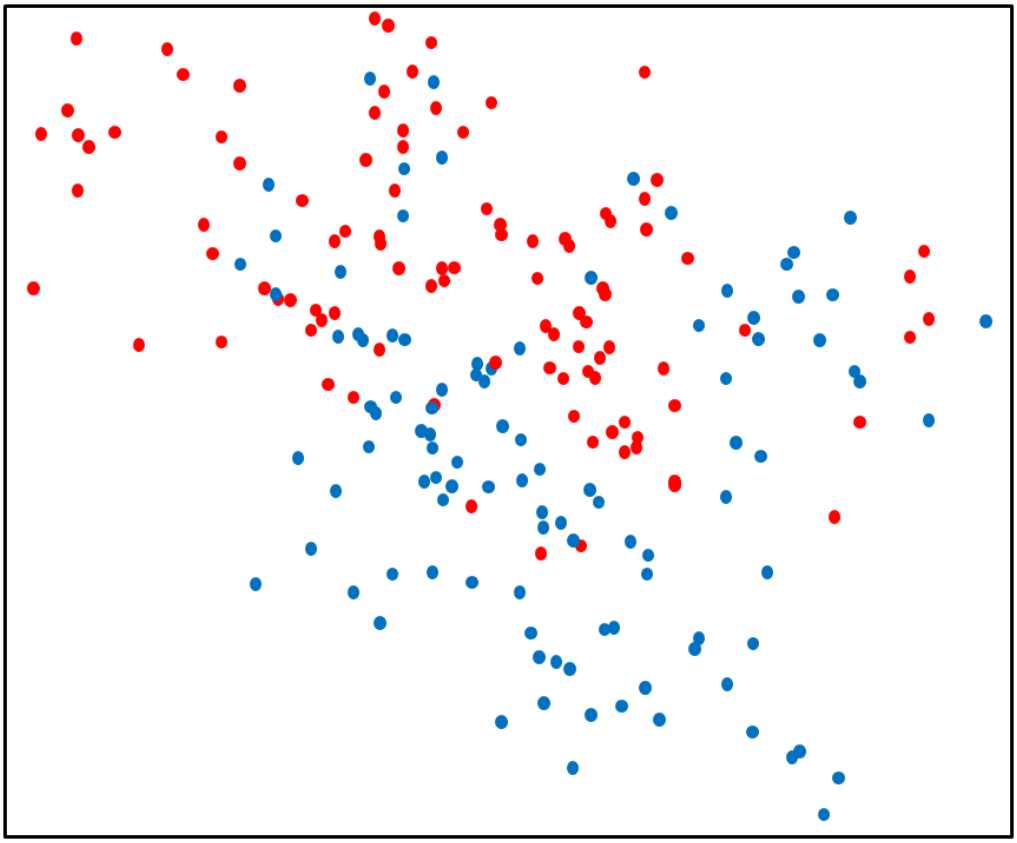
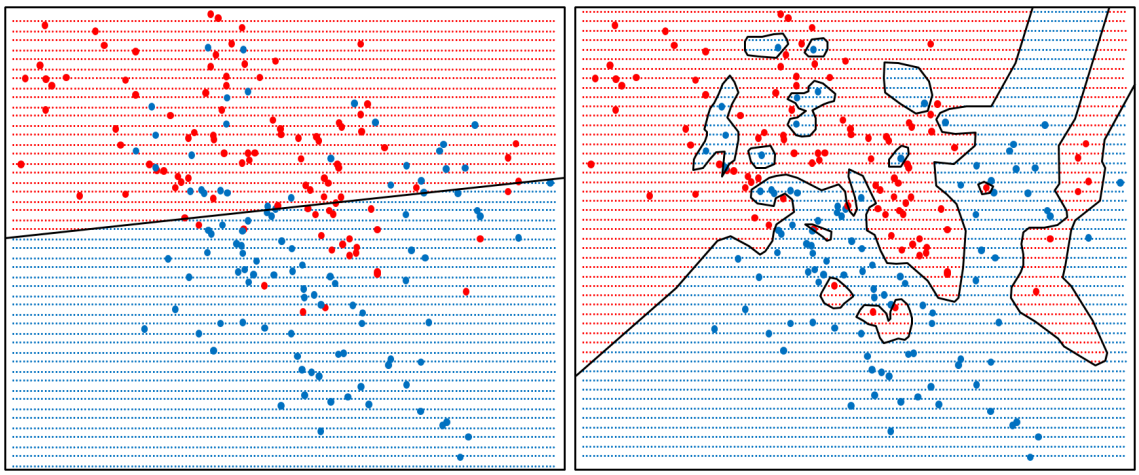


Abbildung 4.1: In einer Ebene verteilte, farbige Datenpunkte. Die Position in der Ebene bildet den Input, die Farbe den Output für Schätzmodelle. Die Datenpunkte wurden erzeugt, indem für die blaue Klasse 10 Mittelwerte aus einer bivariaten Gaußverteilung $N((1,0),\boldsymbol{I})$ heraus gezogen worden sind. Analog waren 10 Mittelwerte für die rote Klasse aus einer bivariaten Gaußverteilung $N((0,1),\boldsymbol{I})$ zu erstellen. Für jede Klasse wurden nun 100 Datenpunkte erstellt, indem für jeden Datenpunkt einer der Mittelwerte m_k seiner Klasse mit der Wahrscheinlichkeit 0.1 zu wählen war. Die Position des Datenpunktes ist per Normalverteilung $N(m_k,\boldsymbol{I}/5)$ zu bestimmen. Damit entstand ein Mix aus Gaußschen Clustern für jede Farbklasse.

Um die Unterschiede der beiden Modelle bezüglich des Bias und der Varianz aufzuzeigen, wurden beide Methoden auf dasselbe Beispiel von zwei verschiedenfarbigen, in einer Ebene verteilten Datenpunkten angewandt. Diese Daten sind in Abbildung 4.1 zu sehen. Die Position in der Ebene bildet den Modellinput, die Farbe den Modelloutput. Die Argumentation wurde [HAS08] entnommen.

Bei vorliegender Aufgabe handelt es sich um ein Klassifikationsproblem. Ziel ist es, Bereiche zu definieren in denen alle neuen Punkte eindeutig einer der beiden Farbklassen zuzuordnen sind. Das lineare Modell wird versuchen, eine einfache Trennlinie zu errichten, welche anhand der vorliegenden Datenpunkte die geringste Fehlerrate aufweist.



(a) Färbung der Ebene mittels linearem Modell. Ein Datenpunkt x unbekannter Farbe wird als „blau“ deklariert, wenn $x^{\mathsf{T}}\boldsymbol{\beta} < 0.5$. Bei Gleichheit liegt eine **Entscheidungsgrenze** vor, bei der das Verhalten des Algorithmus im Zweifel händisch festgelegt werden muss. Im anderen Fall > 0.5 sei der Datenpunkt „rot“.

(b) Färbung der Ebene mittels 1-nearest-neighbor-Modell. Ein Datenpunkt x unbekannter Farbe wird als „blau“ deklariert, wenn $\widehat{Y}(x) < 0.5$. Bei Gleichheit liegt eine **Entscheidungsgrenze** vor, bei der das Verhalten des Algorithmus im Zweifel händisch festgelegt werden muss. Im anderen Fall > 0.5 sei der Datenpunkt „rot“.

Abbildung 4.2: Färbung der Ebene mittels der in den Abschnitten 4.4.1 und 4.4.2

Das knn-Modell sucht für jeden neuen Punkt nach der mehrheitlich auftretenden Farbe seiner k nächsten Nachbarn. Interessant sind vor allem die Bereiche, bei denen die Zuordnung nicht eindeutig ist oder die Entscheidung nur sehr knapp ausfällt. Dies sind die Grenzbereiche. Innerhalb der Entscheidungsgrenzen entsteht ein Bereich, in dem alle Entscheidungen zugunsten einer Farbe ausfallen.

Im Folgenden werden die Lösungen der beiden Modelle vorgestellt. Die Ermittlung der jeweils optimalen Einstellungen erfolgt dabei durch die in Abbildung 4.1 vorgestellten Datenpunkte. Um ein optisches Feedback zu erzeugen, wird ein Gitter an Datenpunkten über die Ebene gelegt. Für jeden Punkt wird anhand des Modells die Färbung bestimmt. Das Vorgehen ist in den Bildbeschreibungen festgehalten.

Es ist offensichtlich, dass das Bias des linearen Modells hoch sein muss. Die Voraussetzung eines linearen Modells wird der wirklichen Datenverteilung aus mehreren Gaußschen Datenclustern also nicht gerecht. Dafür schwankt das Modell nicht in seiner Robustheit, wenn neue Datenpunkte präsentiert werden. Würden komplett neue Datenpunkte erzeugt, so verliefe die neue Entscheidungsgrenze fast analog zur alten. In anderen Worten: Es sind zwar zahlreiche Missklassifikationen auf beiden Seiten gegeben. Bei neuen (auch unbekannten) Datenpunkten wird sich diese Missklassifikationsrate aber kaum ändern. Daher ist die Varianz dieses Modell gering.
Der Missklassifikationsfehler entstammt fast nur der ungünstigen Modellannahme. Wäre die Input-Output-Beziehung ebenfalls linear gewesen, so läge der Bias bei 0. Da die

Varianz bei diesem Modell von vornherein niedrig ist, hätte sich ein optimales Modell ergeben.

Das Bias des 1nn-Modells ist recht niedrig, da dieses Modell für Datencluster wie diese recht gut geeignet ist. Aber auch eine lineare Verteilung der Daten wäre gut modelliert worden. Dies funktioniert, da das Verfahren in der Art seiner Berechnung keinerlei Voraussetzungen an das wahre Modell stellt. Damit sind können alle möglichen Verteilungen bestimmt werden, womit der Bias von Natur aus reduziert ist.
Das Problem liegt jedoch in der hohen Varianz. Wie zu sehen ist, ergibt sich eine sehr unregelmäßige Entscheidungsgrenze, welche teilweise aus winzigen Clustern besteht, die auf einzelnen Datenpunkten beruhen. Würden komplett neue Datenpunkte erzeugt, so ist stark davon auszugehen, dass der Verlauf der Entscheidungsgrenze sich stark von der alten Grenze unterscheidet. In anderen Worten: Die Missklassifikationsrate mag zwar für die Trainingsdaten gering sein. Bei der Anwendung auf Testdaten ist jedoch davon auszugehen, dass sie sehr hoch ist, da sich die Farbzuordnung auf nur einen Datenpunkt bezieht, welcher sich lediglich zufällig an dieser Position befindet. Vor allem Grenzbereiche werden damit sehr schlecht abgebildet.

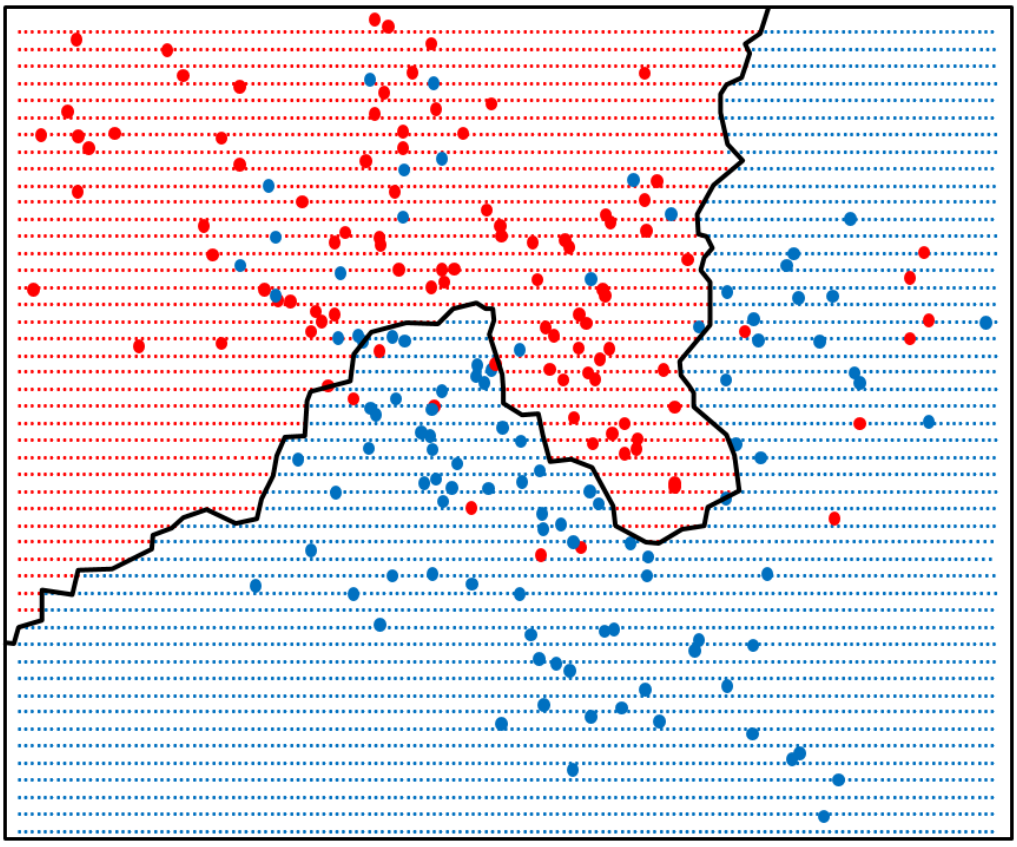


Abbildung 4.3: Färbung der Ebene mittels 1-nearest-neighbor-Modell. Ein Datenpunkt x unbekannter Farbe wird als „blau“ deklariert, wenn $\widehat{Y}(x) < 0.5$. Bei Gleichheit liegt eine **Entscheidungsgrenze** vor, bei der das Verhalten des Algorithmus im Zweifel händisch festgelegt werden muss. Im anderen Fall > 0.5 sei der Datenpunkt „rot“.

Einen guten Kompromiss liefert in diesem Fall das in Abbildung 4.3 gezeigte 15-nearest-neighbor-Modell. Durch die Einbeziehung einer größeren Nachbarschaft geht die Tendenz einer größeren Menge von Punkten ein. Das senkt die Varianz erheblich. Gleichzeitig wird das Bias etwas erhöht, da die Modellkomplexität durch die größere Nachbarschaft gestiegen ist. Da das Modell aber ohnehin einen geringen Bias besitzt, fällt die Erhöhung nicht so sehr ins Gewicht wie die Verringerung der Varianz.

4.4.4 Kreuzvalidierung

Wie im letzten Abschnitt zu sehen war, ist die Wahl einer ausreichenden Modellkomplexität neben der eigentlichen Modellwahl von entscheidender Bedeutung. Die Wahl eines passenden Verfahrens kann aufgrund der Analyse der Daten im Rahmen des sogenannten **Data Minings** erfolgen. In dieser Arbeit wird dieser Abschnitt umgangen, indem verschiedene Modelle angewandt werden, die in der Fachliteratur dieses Bereiches häufig auftreten.
Die Wahl der Modellkomplexität ist dagegen wesentlich einfacher zu gestalten. Sie erfolgt im Rahmen einer sogenannten einfachen k-fachen Kreuzvalidierung oder auch nur **Kreuzvalidierung** genannt („cross validation“). Dazu müssen die Daten in drei Mengen aufgeteilt werden: Der bereits bekannten **Trainingsmenge**, der **Testmenge** und der **Validierungsmenge**.

Die Aufteilung geschieht folgendermaßen: Zuerst werden alle Daten in $k+1$ gleich große Mengen aufgeteilt. Eine beliebige Menge wird als Validierungsmenge ausgewählt. Danach erfolgt die Wahl der Testmenge aus einer der verbliebenen k Mengen.
Mittels der Trainingsmenge ist das Modell zu optimieren und auf der Testmenge der Fehler zu bestimmen. Dieser Vorgang wird k-mal wiederholt. Dabei ist die Testmenge jedesmal neu aus den k Mengen zu ziehen, sodass am Ende der k Durchläufe jede Menge einmal Testmenge war. der Gesamtfehler wird als Durchschnitt der Einzelfehler berechnet. Ist er dem Anwender zu groß, so ist das Modell in seiner Komplexität oder anderen Einstellungen zu ändern. Danach erfolgt eine erneute k-fache Anpassung auf den Trainingsdaten und schließlich die Fehlerbestimmung auf den Testdaten.
Dabei gilt: Je höher die Komplexität, desto geringer ist das Bias, da mehr Datenpunkte erfasst werden. Gleichzeitig vergrößert diese Erfassung die Schwankung in den Daten, was sich als erhöhte Varianz auswirkt. Das Ziel ist also eine Komplexität, welche den Gesamtfehler auf den Testdaten minimiert. Abbildung 4.4 gibt einen Überblick dazu, wie sich die Fehler in den Trainings- und Testdaten bezüglich der Komplexität üblicherweise verhalten.
Anschließend erfolgt die Eingabe der Validierungsdaten in das optimierte Modell, um den abschließenden Fehler festzustellen. Diese Daten haben den Vorteil, dass sie noch nicht verwendet wurden und das Modell damit nicht auf sie angepasst werden konnte. Weicht der Fehler dieser Daten nicht signifikant vom Testfehler ab, so wurde eine gute Einstellung des Modells gefunden. Bei großer Abweichung kann es sein, dass auch

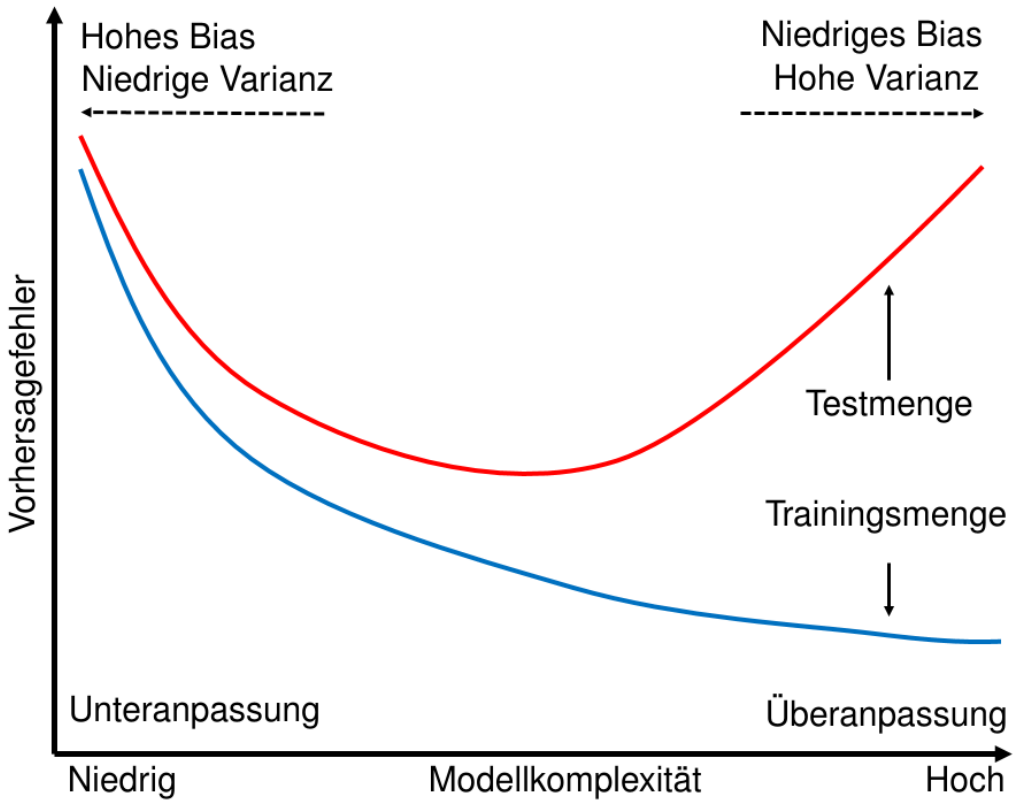


Abbildung 4.4: Darstellung der Vorhersagefehler eines Modells für Trainings- und Testdaten bei Änderung der Modellkomplexität. Die rote Linie steht für den Fehler der Testdaten, die blaue für den Fehler der Trainingsdaten. Niedrige Modellkomplexität führt zu hohem Bias und geringer Varianz, was in Unteranpassung resultiert. Hohe Modellkomplexität führt zu geringem Bias und hoher Varianz, was in Überanpassung resultiert.

eine Überanpassung auf die Testmenge stattgefunden hat, die Daten waren nicht gut ausgewählt oder nicht repräsentativ. Auf jeden Fall sollten alle Mengen neu erstellt werden. Ist auch hierbei eine deutliche Abweichung des Test- vom Validierungsfehler zu verzeichnen, sollte ein Modellwechsel in Betracht gezogen werden.

4.4.5 Allgemeine Einflüsse und Probleme von Schätzmodellen

Neben der bereits vorgestellten Komplexität gibt es viele weitere Einflüsse, welche sich auf den Bias und die Varianz eines Schätzmodells auswirken. Einige wenige davon sollen mit ihren zum Teil großen Auswirkungen in diesem Abschnitt kurz vorgestellt werden.

Es ist leicht vorstellbar, dass eine mangelnde Anzahl an Trainingspunkten zu schlechteren Ergebnissen der Schätzung führt. So muss beim k-nearest-neighbor-Modell auf weit entfernte Datenpunkte zurückgreifen, um die Nachbarschaft der k nächsten Punkte zu erstellen. Ergo liegen die Outputs metrisch gesehen nicht nah beieinander, was zu

Fehlern im Modell führt. Das lineare Modell dagegen benötigt aufgrund seiner Globalität weniger Datenpunkte für eine hinreichend genaue Schätzung als ein lokal orientiertes Verfahren wie das knn. Bei der Modellwahl sollte auf einen solchen Umstand geachtet werden. Manchmal führen auch strukturierte Modellannahmen zu einer effizienteren Nutzung.
Eine große Trainingsmenge minimiert also das Risiko einer Fehlinterpretation der Daten. Bei kleinem N ist neben der Wahl eines neuen Modells die Beschränkung der Parameter wichtig. Weniger Parameter benötigen weniger Daten zur Anpassung. Damit wird die Schwierigkeit der Problemlösung auf die Schwierigkeit der Beschränkungsauswahl verschoben. Üblicherweise ist damit eine Beschränkung der Komplexität gemeint. Im polynomialen Sinn wird, wie bereits oben erwähnt, das Verfahren globalisiert.

Auch eine Verringerung der Dimension der Inputdaten kann hilfreich sein. Denn vor allem bei lokal orientierten Verfahren greift bei einer großen Anzahl von Inputparametern der „**Fluch der Dimensionen**“. Im Falle eines hochdimensionalen Inputs tritt oben beschriebenes Problem der vergrößerten Nachbarschaftsreichweite deutlich verstärkt zu Tage. Zur Veranschaulichung sei erneut das k-nearest-neighbor-Modell herangezogen. Betrachtet sei das knn-Verfahren in einem p-dimensionaler Hyperwürfel (siehe Abbildung 4.5).

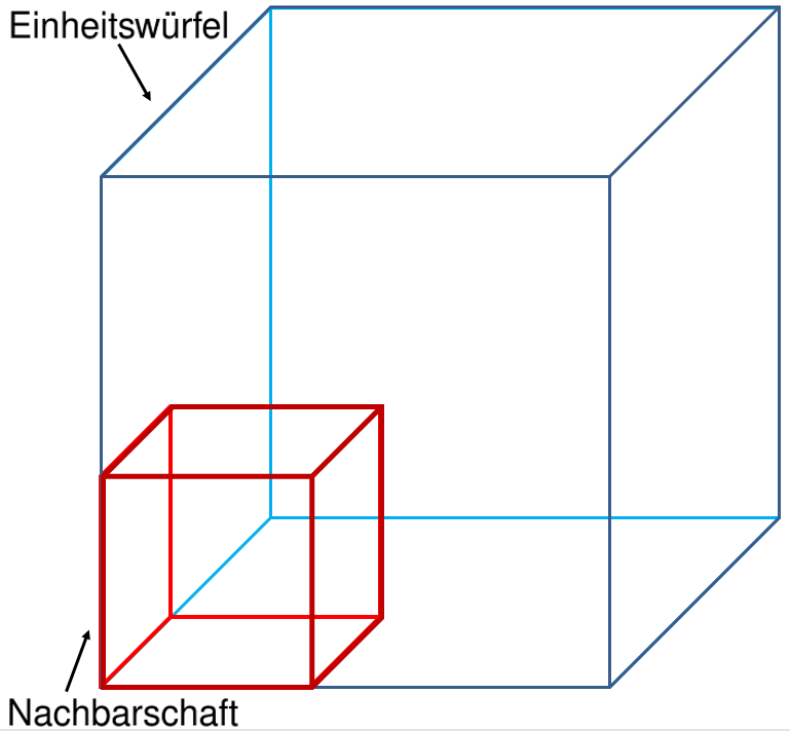


Abbildung 4.5: Einheitshyperwürfel für den dreidimensionalen Fall. Der rote Würfel gibt die Nachbarschaft um einen Datenpunkt an.

Dazu sei ein Datenpunkt betrachtet, dessen Output zu bestimmen ist. Seine benötigte Reichweite, um eine hyperkubische Nachbarschaft der Mächtigkeit k zu erreichen, sei rot gekennzeichnet. Um einen Bruchteil von r Prozent der Gesamtdaten im roten Würfel zu vereinen, muss seine Kantenlänge in Abhängigkeit von Dimension p und Bruchteil r bestimmt werden. Die erwartete Kantenlänge e lautet $e_p(r) = r^{1/p}$. Beispielsweise ergibt sich für ein Prozent der Daten bei einem zehndimensionalen Input eine Kantenlänge von $e_{10}(0.01) = 0.63$ und für zehn Prozent der Daten $e_{10}(0.1) = 0.80$. Eine solche Nachbarschaft kann nicht länger als „lokal“ angesehen werden. Ein geringeres r entspricht einem kleinerem k und führt zu erhöhter Varianz.
Ein weiteres Problem hochdimensionaler Daten ist die Nähe zu einer Kante des Datenraums. Zur Verdeutlichung seien die Daten in einem p-dimensionalen Einheitsball um den Ursprung gleichmäßig verteilt. Die nearest-neighbor-Schätzung soll nun für den Ursprung durchgeführt werden. Die mittlere Distanz vom Ursprung zum nächsten Datenpunkt ist über den Ausdruck

$$d(p,N) = \left(1 - \frac{1}{2}^{\frac{1}{N}}\right)^{\frac{1}{p}} \tag{4.13}$$

bestimmbar [HAS08]. Eine wesentlich kompliziertere Darstellung existiert für den Mittelwert. Für $N = 500$ und $p = 10$ ergibt sich ein Abstand von $d(10,500) \approx 0.52$, der über die Hälfte näher an der Datengrenze liegt als am Zentrum und damit am zu schätzenden Datenpunkt. Damit liegen die meisten Datenpunkte einer Nachbarschaft näher an der Grenze als an irgendeinem anderen Punkt. Das Problem dabei ist, das eine genaue Berechnung gerade an den Grenzen wesentlich schwerer ist als im Inneren. Statt Interpolationen sind dort Extrapolationen von benachbarten Mengenpunkten nötig.
Ein weiteres Problem liefert die benötigte Datendichte. Einen höhere Inputdimension benötigt aufgrund eines größeren Raumes logischerweise mehr Daten, um eine annähernd ähnlich ausreichende Datenmenge für Schätzungen wie im niedrigdimensionalen Fall zu erhalten. Diese Datendichte ist proportional zu $N^{1/p}$. Stellt also $N_1 = 100$ eine ausreichende Datendichte im eindimensionalen Fall dar, so werden im zehndimensionalen Fall $N_{10} = 100^{10}$ Datenpunkte benötigt, um eine vergleichbare Datendichte zu erhalten.

Im hochdimensionalen Fall können jedoch allgemeinere Probleme auftreten, welche auch Verfahren globaler Natur betreffen. So zum Beispiel wächst die Komplexität einer Schätzfunktion exponentiell mit der Dimension. Dies ist prinzipiell gut, da mehr Datenpunkte erfasst werden, was den Bias verringert. Leider wachsen damit auch die Schwankungen in den Daten, was die Varianz erhöht. Bei einem exponentiellen Wachstum droht damit die Gefahr einer Überanpassung. Um eine annähernd gleichbleibende Genauigkeit der Schätzung wie im niedrigdimensionalen Fall beizubehalten, muss die Datenmenge exponentiell mit der Dimension wachsen.
Weiterhin bleibt zwar die Konvergenz eines Verfahrens bestehen, die Konvergenzrate

sinkt jedoch mit der Dimensionshöhe. Dies bedeutet ein Mehr an Rechenzeit.

Im Allgemeinen scheint es, dass weniger komplexe Methoden oder auch globale Verfahren wie das lineare Modell vergleichsweise mehr oder weniger immun gegenüber hohen Dimensionen erscheinen. Wie in [HAS08] in Abschnitt 2.5 nachzulesen ist, existiert eine Darstellung des erwarteten Vorhersagefehlers als $\sigma^2(p/N) + \sigma^2$. Der Fehler wächst also nur linear in der Dimension. Mit einer ausreichend großen Datenmenge ist er zu vernachlässigen.
Weitere Möglichkeiten um einige der auftretenden Probleme zu reduzieren sind neben der bereits erwähnten Reduktion der Modellparameter (Verringerung der Komplexität) die Reduktion der Dimension und die sogenannte Feature Selection. Per Einsatz von Analysemethoden ist zu eruieren, welche Inputparameter einen geringen Einfluss auf den Output besitzen und daher entfernt werden können. Das Ergebnis ist ein Modell mit erhöhtem Bias aber reduzierter Varianz. Im Gegensatz dazu erhöht die Erweiterung der Daten um wichtige Inputparameter die Varianz und senkt den Bias. Ein wohlüberlegter Austausch von unwichtigen gegen wichtige Daten ist im Zweifelsfall die beste Variante.
Wie bereits mehrfach gesagt wurde, ist ein möglichst großer Satz an Trainingsdaten äußerst hilfreich. Es kann nie zu viele Daten geben. Mithilfe der Feature Selection können eventuell Datensätze genutzt werden, die vorher aufgrund von Unvollständigkeit auszuschließen waren. Derlei Datensätze existieren in dieser Arbeit.
Im Allgemeinen wird zu Modellen mit Mischverteilungen und/oder zu einer Kombination unterschiedlicher Lernmethoden geraten. Erwähnt sei dabei der Begriff des Boostings und des Baggings. Siehe dazu [HAS08].

Das in dieser Arbeit behandelte Problem besteht zeitweise aus mehr als 20 Inputvariablen. Damit liegt es nahe, globale statt lokale Methoden zu verwenden. Für eine für alle Wege und/oder Routen gültige Variable stehen pro verwendetem Jahr ca. 1.5 Millionen Routen und damit Datensätze zur Verfügung. Trotz der hohen Dimension erscheint diese Datenmenge ausreichend zu sein. Inwiefern die Aussagen dieses Absatzes zutreffen, ist in der Auswertung zu sehen.

5 Die Vorhersagemodelle

Dieses Kapitel widmet sich den Schätzmodellen zur Vorhersage der Aufteilung der Passagiere der DD auf unterschiedliche Routen. Es gibt 5 Modelle, die mehr oder weniger aufeinander aufbauen.

5.1 Modell 1: Naiver Vergleich

Diese Methode verfolgt den naivsten denkbaren Ansatz. Die Passagierzahlen aus den DPD eines beliebigen Monats werden als die Passagierzahlen des zu schätzenden Monats erklärt. Dazu werden die relativen Häufigkeiten jeder Route bezüglich ihrer Inanspruchnahme durch die Passagiere ermittelt. Diese werden als die geltenden relativen Häufigkeiten der Routen des zu schätzenden Monats angesehen. Günstig kann hierbei der Vormonat oder derselbe Monat des Vorjahres sein.
Mit dieser Methode soll eine erste Meinung und Abschätzung über die prinzipielle Güte der Daten, der Monate und deren jeweilige Abweichung im Bezug auf andere Monate aufgestellt werden.
Der einmalige Durchlauf der Methode geht recht schnell vonstatten. Somit ist es möglich, sämtliche Monate in kurzer Zeit gegenüberzustellen.

5.1.1 Variablen

Für das Modell werden in Tabelle 5.1einige Variablen eingeführt, die für diesen Abschnitt gelten.

Variable	Beschreibung
t_{fut}	Oberer Index, der den zu schätzenden Zeitpunkt kennzeichnet
t	Oberer Index, der einen Zeitpunkt kennzeichnet
J	Anzahl der Routen eines Weges
$j = 1,...,J$	Route eines Weges
y_j	Anzahl der Passagiere auf Route j entspricht den Werten der DPD
$Y = \sum_{j=1}^{J} y_j$	Gesamtzahl der Passagiere eines Weges; entspricht den Werten der DD

Tabelle 5.1: Variablen dieses Abschnittes

5.1.2 Das Modell

Es ist zu beachten, dass die Bearbeitung für jeden Weg einzeln erfolgt. Zur Verbesserung der Lesbarkeit wird daher auf einen Wegindex verzichtet.
Die Auswahlwahrscheinlichkeit für Route j zum Zeitpunkt t entspricht der relativen Häufigkeit der Route j zum Zeitpunkt t

$$\pi_j^{(t)} = \frac{y_j^{(t)}}{Y^{(t)}}. \tag{5.1}$$

Der Vektor $\boldsymbol{\pi}^{(t)} = \begin{pmatrix} \pi_1^{(t)} \\ \vdots \\ \pi_J^{(t)} \end{pmatrix}$ beschreibt die relativen Wahrscheinlichkeiten der Aufteilung der Passagiere eines Weges zum Zeitpunkt t auf die verschiedenen Routen.

5.1.3 Die geschätzte Passagierzahl pro Route

Mittels einer skalaren Multiplikation mit der aus den DD gegebenen Passagierzahl $Y^{(t_{\text{fut}})}$ eines Weges und den relativen Häufigkeiten der Passagiere auf den Routen zu einem älteren Zeitpunkt t ergibt sich die geschätzte absolute Verteilung der Passagiere auf die Routen

$$\begin{pmatrix} \widehat{y}_1^{(t_{\text{fut}})} \\ \vdots \\ \widehat{y}_J^{(t_{\text{fut}})} \end{pmatrix} = Y^{(t_{\text{fut}})} \boldsymbol{\pi}^{(t)}. \tag{5.2}$$

In den nachfolgenden Modellen werden effizientere Methoden zur Ermittlung von $\boldsymbol{\pi}^{(t)}$ behandelt. Die Bestimmung der absoluten Verteilung der Passagiere erfolgt anschließend mittels Gleichung (5.2).

5.1.4 Bemerkungen

Problem: Routen des Zeitpunktes t_{fut}, die nicht während t vorkommen, werden nicht erkannt. Damit addieren sich die Wahrscheinlichkeiten der nicht erkannten Routen auch nicht zu den erkannten Routen. Dies führt zu einer Abweichung unbekannten Ausmaßes.
Eine Lösung könnte sein, die Routen weiterer Monate mit einzubeziehen. In diesem Falle werden die relativen Häufigkeiten der nicht erkannten Routen zu den π der erkannten Routen addiert und die Wahrscheinlichkeiten auf 1 normiert.

Die Einbeziehung weiterer Routen führt zum linearen Modell des nächsten Abschnitts.

5.2 Modell 2: Lineares Modell mit multiplen Output

Das lineare Modell mit multiplen Output bildet eine Erweiterung des naiven Vergleichs von Abschnitt 5.1. Es vereint zwei wichtige Gesichtspunkte:

1.) Es existieren mehrere **Referenzmonate**. Dabei handelt es sich um die Monate, deren DD und DPD in ein Vorhersagemodell einfließen. Modell 1 besitzt genau einen Referenzmonat. Bei diesem Modell sind es mehrere. Damit fließen mehr Informationen ein.

2.) Aus den Referenzmonaten wird eine Auswahl von Monaten getroffen, deren Routen als die Routen des Vorhersagezeitpunktes t_{fut} gelten können. Die Anzahl der zum Zeitpunkt t_{fut} vorhergesagten und wirklich existierenden Routen wächst. Allerdings auch der Anteil der vorhergesagten und nicht existierenden Routen.
Eine Balancierung der Güte dieses Modells erfolgt über die Anzahl und Art der Referenzmonate sowie über die Anzahl und Art der Auswahl dieser Monate zur Vorhersage der Existenz von Routen für t_{fut}. Es sei erwähnt, dass die **Fehlerfunktion** stets maßgeblich zur Modellgüte beiträgt. Verwendet wurde nur die quadratische Fehlerfunktion. Die Gründe dafür sind in Kapitel 4 dargelegt.

Es sei an dieser Stelle festgehalten, dass Herleitung und Argumentation [AGR02] folgen.

Bemerkung 5.2.1
Die Art der Referenzmonate spaltet sich in zwei Fälle auf: Monatliche und jährliche Zeitpunkte. Für die monatlichen Zeitpunkte spricht ihre Aktualität. Für die jährlichen Zeitpunkte spricht die Resistenz ihrer Daten gegenüber saisonalen Schwankungen (siehe Abschnitt 3.2.1).
Die Art der Monatsauswahl ist sofort klärbar. Es werden die D jüngsten vorhandenen Monate verwendet, da ihre Informationen am aktuellsten sind.
Die Routen des zu schätzenden Monats werden aus allen Routen der D jüngsten Referenzmonate gebildet. Damit soll ausgeschlossen werden, dass Routen auftreten, die nicht mehr aktuell sind.

5.2.1 Variablen

Für das Modell werden in Tabelle 5.2 einige Variablen eingeführt, die für diesen Abschnitt gelten. Die Variablen sind für einen Weg und seine Routen definiert.

Variable	Beschreibung
t_{fut}	Oberer Index, der den zu schätzenden Zeitpunkt kennzeichnet
t	Oberer Index, der einen Zeitpunkt kennzeichnet
N	Anzahl Referenzmonate
J	Anzahl Routen für einen Weg
$\boldsymbol{x}$	$N \times 1$ - Inputvektor, der Gesamtpassagierzahl jeden der N Referenzmonate enthält; gewonnen aus den DD
x	Inputvariable, die aus $\boldsymbol{x}$ bei $N = 1$ entsteht
$\boldsymbol{Y}$	$N \times J$ - Outputmatrix, die für jeden der N Referenzmonate die Aufteilung der Passagiere auf die Routen enthält, gewonnen aus den DPD
$\boldsymbol{y}$	$1 \times J$ - Outputvektor, der aus $\boldsymbol{Y}$ bei $N = 1$ entsteht
$\boldsymbol{\beta}$	$1 \times J$ auf 1 normierter Vektor, der die relativen Häufigkeiten zur Aufteilung der Passagiere auf die Routen enthält
λ	Vorfaktor des Strafterms der Nebenbedingung - je größer λ desto größer die Strafe bei Abweichung vom optimalen Wert der Nebenbedingung
$\mathbf{1}$	$1 \times R$ - Einsvektor

Tabelle 5.2: Variablen dieses Abschnittes

5.2.2 Das Modell

Wie $\boldsymbol{\pi}$ bei Modell 1 (siehe Gleichung (5.1) auf Seite 52) soll es für den zu schätzenden Monat des Zeitpunktes t_{fut} einen Verteilungsschlüssel $\boldsymbol{\beta}$ geben, der die Gesamtpassagiere $x^{(t_{\text{fut}})}$ eines Weges auf die Routen verteilt. Für den zu schätzenden Zeitpunkt $t_{(\text{fut})}$ ($N = 1$) besitzt das Modell folgendes Aussehen für die Routen eines Weges

$$\widehat{\boldsymbol{y}} = x^{(t_{\text{fut}})}\boldsymbol{\beta}. \tag{5.3}$$

Das vorgestellte Verfahren schätzt für jeden Weg ein individuelles $\boldsymbol{\beta}$. Jedes Element des $\boldsymbol{\beta}$-Vektors gibt die relative Häufigkeit derjenigen Passagiere an, welche die Route an der Position des Elementes benutzen.

5.2.3 Schätzung des Modells

Mittels der Referenzmonate kann $\boldsymbol{\beta}$ für jeden Referenzmonat einzeln, in Form der relativen Häufigkeiten, ermittelt werden. Es würden allerdings N verschiedene $\boldsymbol{\beta}$ entstehen.

Gewünscht ist aber ein einziges $\widehat{\boldsymbol{\beta}}$, dass eine möglichst genaue Schätzung für die Referenzmonate liefert. Aus allen bekannten $\boldsymbol{y}$ entsteht folgendes Gleichungssystem

$$\boldsymbol{Y} = \boldsymbol{x}\widehat{\boldsymbol{\beta}}.$$

$\boldsymbol{\beta}$ wird geändert, bis die Abweichungen zu den wirklichen Werten minimal sind. Das gewonnene $\widehat{\boldsymbol{\beta}}$ wird als Verteilungsschlüssel für die Modellgleichung (5.3) genutzt. Eine solche Anpassung des Schätzwertes mittels der Abweichung von bekannten Outputs wird **Supervised Learning**, also **Überwachtes Lernen**, genannt. Siehe dazu Kapitel 4.

Zur Bestimmung eines Wertes für $\widehat{\boldsymbol{\beta}}$ ist die erste Ableitung der Fehlerfunktion nach $\boldsymbol{\beta}$ 0 zu setzen und nach $\boldsymbol{\beta}$ umzustellen. Als zu integrierende Nebenbedingung wird die Normierung von $\boldsymbol{\beta}$ zum Wert 1 verlangt. $\boldsymbol{\beta}$ soll immerhin relative Wahrscheinlichkeiten repräsentieren. Die Nebenbedingung tritt in der Fehlerfunktion als Strafterm mit Vorfaktor λ auf. Damit ist ebenfalls eine Ableitung der Funktion nach λ erforderlich, die es 0 zu setzen und nach λ umzustellen gilt. Die ermittelte Formel wird anschließend in die Formel für $\boldsymbol{\beta}$ eingesetzt.

5.2.4 Herleitung von $\widehat{\boldsymbol{\beta}}$

Anpassungsmodell:

$$\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\beta} \tag{5.4}$$

Zu minimierende Fehlerfunktion mit Nebenbedingung:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\|\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}\|_2^2 \right) \qquad \text{mit } \|\boldsymbol{\beta}\|_1 = 1 \tag{5.5}$$

Herleitung von $\widehat{\boldsymbol{\beta}}$: Minimierungsfunktion mit Lagrangemultiplikator:

$$f(\boldsymbol{\beta}, \lambda) = \sum_{k=1}^{J} \sum_{i=1}^{N} (y_{ik} - x_i \beta_k)^2 - \lambda \left(1 - \sum_{j=1}^{J} \beta_j \right) \tag{5.6}$$

1. Schritt: Ableitung von (5.6) nach $\boldsymbol{\beta}$

$$\frac{\partial f}{\partial \beta_k} = 2\sum_{i=1}^{N} -x_i(y_{ik} - x_i\beta_k) + \lambda = 0.$$

Somit

$$\beta_k = \frac{2\sum\limits_{i=1}^{N} x_i y_{ik} - \lambda}{2\sum\limits_{i=1}^{N} x_i^2}. \tag{5.7}$$

2. Schritt: Ableitung von (5.6) nach λ

$$\frac{\partial f}{\partial \lambda} = -(1 - \sum_{j=1}^{J} \beta_j) = 0.$$

Umstellen nach 1 und Ersetzen von β_j mittels (5.7) ergibt

$$1 = \sum_{k=1}^{J} \beta_k = \frac{\sum\limits_{k=1}^{J}\sum\limits_{i=1}^{N} x_i y_{ik} - \frac{1}{2}J\lambda}{\sum\limits_{i=1}^{N} x_i^2}$$

$$\sum_{i=1}^{N} x_i^2 = \sum_{k=1}^{J}\sum_{i=1}^{N} x_i y_{ik} - \frac{1}{2}J\lambda.$$

Umstellen liefert

$$\lambda = \frac{2}{J}\left(\sum_{k=1}^{J}\sum_{i=1}^{N} x_i y_{ik} - \sum_{i=1}^{N} x_i^2\right). \tag{5.8}$$

3. Schritt: Ersetze λ in (5.6)

$$\begin{aligned}\beta_k &= \frac{\sum\limits_{i=1}^{N} x_i y_{ik}}{\sum\limits_{i=1}^{N} x_i^2} - \frac{\lambda}{2\sum\limits_{i=1}^{N} x_i^2} \\ &= \frac{\sum\limits_{i=1}^{N} x_i y_{ik}}{\sum\limits_{i=1}^{N} x_i^2} - \frac{\sum\limits_{j=1}^{J}\sum\limits_{i=1}^{N} x_i y_{ij}}{J\sum\limits_{i=1}^{N} x_i^2} + \frac{\sum\limits_{i=1}^{N} x_i^2}{J\sum\limits_{i=1}^{N} x_i^2}.\end{aligned} \tag{5.9}$$

4. Schritt: Zusammenfassung Schritt 3 in Vektorschreibweise

$$\widehat{\boldsymbol{\beta}} = \frac{1}{\|\boldsymbol{x}\|^2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y} + \mathbf{1}\left(\frac{\|\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y}\|_1}{J\|\boldsymbol{x}\|^2} + \frac{1}{J}\right).$$

Das Vorhersagemodell für den zu schätzenden Monat zum Zeitpunkt t_{fut} lautet damit

$$\widehat{\boldsymbol{y}} = x^{(t_{\text{fut}})}\widehat{\boldsymbol{\beta}}$$

mit

$$\widehat{\boldsymbol{\beta}} = \frac{1}{\|\boldsymbol{x}\|^2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y} + \mathbf{1}\left(\frac{\|\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y}\|_1}{J\|\boldsymbol{x}\|^2} + \frac{1}{J}\right). \tag{5.10}$$

5.2.5 Vor- und Nachteile

Ein Vorteil dieses Verfahrens ist die theoretisch hohe Robustheit gegen Ausreißer. Die Existenz von Ausreißern in den Inputdaten beeinflusst das Ergebnis in nur geringem Ausmaß, wie in Kapitel 4 ausführlich gezeigt wurde. Bei der Verwendung von lediglich einem Referenzmonat wie bei Modell 1 ist es möglich, dass eine Route als stark beflogen erscheint, obwohl sie sonst selten nachgefragt ist und umgekehrt. Eine allein darauf basierende Vorhersage neigt zu starken Abweichungen des Outputs im Vergleich mit der Realität. Durch die Betrachtung der Route über mehrere Zeitpunkte hinweg geht die allgemeine Tendenz der Benutzung dieser Route in das Modell ein, sodass die Auswirkung einzelner abweichender Werte abgeschwächt wird.
Weiterhin besticht dieses Verfahrens durch seine Einfachheit und die daraus resultierende Geschwindigkeit. Es existiert eine explizite Formel von $\widehat{\boldsymbol{\beta}}$, sodass kein Zeitaufwand für die Berechnung von Näherungslösungen entsteht. Es eignet sich somit für mehrmalige Wiederholungen, die zum Beispiel bei größer angelegten Tests auftreten. Siehe dazu die untenstehende Bemerkung 5.2.2.
Einen weiteren Vorteil bietet die Erweiterbarkeit. Aktuell besteht der Input lediglich aus der Passagierzahl. Es wären weitere Inputinformationen denkbar, wie die Reisezeit oder die Entfernung. Natürlich würde sich die Fehlerfunktion (5.5) leicht abändern, aus $x_i\widehat{\beta}_k$ würde $\sum_{r=1}^{\text{Anzahl Inputs}} x_{ir}\widehat{\beta}_{kr}$ und die Summe der $\widehat{\boldsymbol{\beta}}$ - Elemente einer Route müsste positiv sein. Es kämen also weitere Nebenbedingungen hinzu. Diese Erweiterung ist mit Methode 4 in Abschnitt 5.4 gelöst. Sie ist zwar eine wesentlich langsamere Näherungsmethode, beinhaltet aber eine Vielzahl von Inputinformationen, deren Güte/Nützlichkeit sich besser analysieren lässt.

Der Hauptnachteil des Modells, welcher eine Anwendung in der Praxis fragwürdig macht, ist, dass ein $\boldsymbol{\beta}$ eine Mächtigkeit von bis zu 70 Elementen erreicht. Dem gegenüber stehen bei jährlichen Referenzzeitpunkten ca. 12 Referenzzeitpunkte zurück bis ins Jahr

2002, mit deren Hilfe $\boldsymbol{\beta}$ angepasst werden kann. Durch die Betrachtung monatlicher Referenzzeitpunkte kann diese Zahl auf knapp 150 vergrößert werden. Die Qualität der älteren Daten ist in diesem Fall jedoch umstritten. Somit liegt die Anzahl der zu anzupassenden Parameter potentiell weit höher als die Anzahl der Datenpunkte, die zur Schätzung herangezogen werden können. Die Qualität der erreichten Anpassung ist somit von zweifelhafter Natur. Siehe dazu Kapitel 4. Zur Lösung dieses Problems wird das nächste Modell eingeführt, welches eine ausreichende Anzahl an Daten erzeugen soll.
Der nächste Nachteil dieses Modells liegt darin, dass es linear und damit recht starr ist. Es liefert nur für (annähernd) lineare Probleme Lösungen mit großer Güte. Für komplexere Probleme sollten entsprechend angepasste und dadurch komplexere Lösungsmodelle verwendet werden. Diese angepassten Modelle bieten allerdings ebenso Risiken, siehe dazu Kapitel 4.

Bemerkung 5.2.2
Der Nutzen des Modells für diese Arbeit besteht in der Feststellung der optimalen Anzahlen für die Referenzmonate und die Auswahlmonate. Und der Gegenüberstellung der Güte bei der Verwendung der monatlichen und jährlichen Zeitpunkte. Die gewonnenen Kennzahlen werden bei den späteren Modellen in Abschnitt 5.4 und 5.5 verwendet und sind nicht erst durch langwieriges Testen zu ermitteln.

5.3 Modell 3: Lineares Bootstrap Modell mit multiplen Output

Das lineare Bootstrap Modell mit multiplen Output ist eine Erweiterung von Modell 2 (siehe Abschnitt 5.2). Statt vieler, weit zurückreichender Referenzmonate werden einige Referenzmonate der jüngeren Vergangenheit mehrmals verwendet. Sollte bei Modell 2 festzustellen sein, dass jährliche DD und DPD von Nöten sind, um der Saisonbereinigung zu entgehen, tritt das bereits in Abschnitt 5.2.5 beschriebene Problem auf: Es herrscht ein Mangel an ausreichend aktuellen Datensätzen. Zwar entwickelt sich der Flugverkehr nicht in dem Maße wie zum Beispiel die Computertechnik, allerdings ist davon auszugehen, dass zehn Jahre alte Flugpläne von mehr oder weniger geringer Bedeutung sind. Doch selbst mit Hilfe dieser Daten ist ihre Anzahl nicht groß genug. Es folgt, dass nur eine kleine Menge an aktuellen Datensätzen existiert. Dieses Problem wird mit Einsatz des Bootstrap behoben. Für jeden Weg werden die Anzahlen der Passagiere aus den DPD auf N verschiedene Datensätze aufgesplittet. Damit wird die nötige Anzahl an Daten erzeugt. Danach wird das $\boldsymbol{\beta}$ analog zu Modell 2 ermittelt. Zur Verifizierung des entstandenen $\boldsymbol{\beta}$ wird dieser Vorgang B-mal wiederholt und alle $\boldsymbol{\beta}$ gemittelt. Der nachfolgend beschriebene Algorithmus ist wie bei Modell 2 für einen Weg beschrieben.

5.3.1 Variablen

Die Variablen dieses Abschnitts sind in Tabelle 5.3 aufgeführt.

Variable	Beschreibung
P	Anzahl der Referenzmonate
B	Anzahl Wiederholungen
$b = 1, ..., B$	Eine spezielle Bootstrap-Wiederholung
N	Anzahl Datensätze
J	Anzahl Routen
$\boldsymbol{x}$	$(N \cdot P) \times 1$ - Inputvektor, der die Gesamtpassagierzahl jedes der N Datensätze für jeden der P Referenzmonate enthält; gewonnen aus den DD
$\boldsymbol{Y}$	$(N \cdot P) \times J$ - Outputmatrix die für jeden der $N \cdot P$ Datensätze die Aufteilung der Passagiere auf die Routen enthält, die in den P Referenzmonaten auftreten; gewonnen aus den DPD
$\boldsymbol{y}$	$1 \times J$ - Vektor, bei dem ein Element die Anzahl der Passagiere einer bestimmten Route repräsentiert
$\boldsymbol{Y}^{(b)}$	Outputmatrix der Wiederholung b
$\boldsymbol{\beta}$	$1 \times J$ - auf 1 normierter Vektor, der die relativen Häufigkeiten zur Aufteilung der Passagiere auf die Routen enthält
$\boldsymbol{\beta}^{(b)}$	Das $\boldsymbol{\beta}$ der Wiederholung b
λ	Vorfaktor des Strafterms der Nebenbedingung; je größer λ, desto größer die Strafe bei Abweichung vom optimalen Wert der Nebenbedingung
$\mathbf{1}$	$1 \times J$ - Einsvektor

Tabelle 5.3: Variablen dieses Abschnittes

5.3.2 Die Erzeugung der Datensätze

Die N Datensätze einer Wiederholung b werden nach folgendem Vorgehen erzeugt. Die eingeführten Bezeichnungen gelten nur für diesen einzelnen Abschnitt:
Betrachtet sei ein Weg zu einem einzelnen Zeitpunkt. Die Anzahl aller Passagiere wird mit n beschrieben. Mit p_j soll die relative Häufigkeit der Passagiere auf Route j beschrieben sein. Sei $n_{j,i}$ die Anzahl der Passagiere auf Route j des i-ten Datensatzes ($i = 1, ..., N$). Dann soll zur Aufteilung der Passagiere auf die Routen der Datensätze folgende Verteilung für den i-ten Datensatz gelten

$$(n_{1,i}) \sim \text{mult}\left(\frac{n}{N}, p_1, ..., p_J\right).$$

Damit sind die Passagiere aufgeteilt. Es sei der Hinweis gegeben, dass die implementatorische Umsetzung sich an diesem Vorgehen orientiert, wegen des letzten Schrittes aber nach einer leicht abgewandelten Vorgehensweise arbeitet.

5.3.3 Das Modell

Das Modell wird zunächst für einen Referenzmonat beschrieben.
Dieser wird in N Datensätze zerlegt. Dies geschieht, indem die Passagiere einer Route, wie in Abschnitt 5.3.2 beschrieben, zufällig gleichverteilt mit Zurücklegen auf die N Datensätze aufgeteilt werden. Diese N Datensätze sind wie bei Modell 2 in Abschnitt 5.2 als N verschiedene Referenzmonate zu behandeln. Damit erfolgt das weitere Vorgehen analog zur Schätzung von $\boldsymbol{\beta}$ aus Modell 2.
Nach Ermittlung von $\widehat{\boldsymbol{\beta}}^{(1)}$ wird die Zerlegung in die Datensätze und die Berechnung des resultierenden $\widehat{\boldsymbol{\beta}}^{(b)}$ $B-1$-mal wiederholt. Anschließend werden die $\widehat{\boldsymbol{\beta}}^{(b)}$ gemittelt. Das Resultat ist folgendes Modell

$$\widehat{\boldsymbol{y}} = x^{(t_{\text{fut}})} \left(\frac{1}{B} \sum_{b=1}^{B} \widehat{\boldsymbol{\beta}}^{(b)} \right). \tag{5.11}$$

Damit ergibt sich für jeden Weg ein individuelles $\widehat{\boldsymbol{\beta}}^{(b)}$, das auf den zu schätzenden Monat angewendet werden kann. Jedes Element des $\widehat{\boldsymbol{\beta}}^{(b)}$-Vektors gibt die relative Häufigkeit der Passagiere an, welche die Route an der Position des Elementes benutzen.

Um das Modell auf mehrere Referenzmonate auszuweiten muss N in den Formeln lediglich durch den Faktor $N \cdot P$ ersetzt werden. Praktisch wird für jeden Monat eine einzelne Outputmatrix $\boldsymbol{Y}$ und ein einzelner Inputvektor $\boldsymbol{x}$ erzeugt, die so konkateniert werden, dass sie der in Abschnitt 5.3.1 dargestellten Form entsprechen. Sollte eine Route in einem Monat nicht auftreten, so sind Nulleinträge einzusetzen. Die zeitliche Reihenfolge der Daten ist unerheblich.

5.3.4 Schätzung des Modells

Anpassungsmodell:

$$\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\beta} \tag{5.12}$$

Mit $\boldsymbol{\beta} = \frac{1}{B} \sum_{b=1}^{B} \boldsymbol{\beta}^{(b)}$

Zu minimierende Fehlerfunktion mit Nebenbedingung

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\|\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}\|_2^2 \right) \quad \text{mit } \|\boldsymbol{\beta}\|_1 = 1. \tag{5.13}$$

Herleitung von $\widehat{\boldsymbol{\beta}}^{(b)}$:
Minimierungsfunktion mit Lagrangemultiplikator

$$f(\boldsymbol{\beta}^{(b)},\lambda) = \sum_{k=1}^{J}\sum_{i=1}^{NP}(y_{ik}^{(b)} - x_i\beta_k^{(b)})^2 - \lambda(1 - \sum_{j=1}^{J}\beta_j^{(b)}). \tag{5.14}$$

1. Schritt: Ableitung von (5.13) nach $\beta_k^{(b)}$

$$\frac{\partial f}{\partial \beta_k^{(b)}} = 2\sum_{i=1}^{NP} -x_i(y_{ik}^{(b)} - x_i\beta_k^{(b)}) + \lambda = 0$$

nach $\beta_k^{(b)}$ umstellen:

$$\beta_k^{(b)} = \frac{2\sum\limits_{i=1}^{NP} x_i y_{ik}^{(b)} - \lambda}{2\sum\limits_{i=1}^{NP} x_i^2}. \tag{5.15}$$

2. Schritt: Ableitung von (5.13) nach λ

$$\frac{\partial f}{\partial \lambda} = -(1 - \sum_{j=1}^{J}\beta_j^{(b)}) = 0.$$

Umstellen nach 1 und Ersetzen von $\beta_j^{(b)}$ mittels (5.15) ergibt:

$$1 = \sum_{k=1}^{J}\beta_k^{(b)} = \frac{\sum\limits_{k=1}^{J}\sum\limits_{i=1}^{NP} x_i y_{ik}^{(b)} - \frac{1}{2}J\lambda}{\sum\limits_{i=1}^{NP} x_i^2} \Longrightarrow \sum_{i=1}^{NP} x_i^2 = \sum_{k=1}^{J}\sum_{i=1}^{NP} x_i y_{ik}^{(b)} - \frac{1}{2}J\lambda.$$

Umstellen nach λ liefert

$$\lambda = \frac{2}{R}\left(\sum_{k=1}^{J}\sum_{i=1}^{NP} x_i y_{ik}^{(b)} - \sum_{i=1}^{NP} x_i^2\right) \tag{5.16}$$

3. Schritt: Ersetze λ in (5.15)

$$\begin{aligned}\beta_k^{(b)} &= \frac{\sum\limits_{i=1}^{NP} x_i y_{ik}^{(b)}}{\sum\limits_{i=1}^{NP} x_i^2} - \frac{\lambda}{2\sum\limits_{i=1}^{NP} x_i^2} \\ &= \frac{\sum\limits_{i=1}^{NP} x_i y_{ik}^{(b)}}{\sum\limits_{i=1}^{NP} x_i^2} - \frac{\sum\limits_{j=1}^{J}\sum\limits_{i=1}^{NP} x_i y_{ij}^{(b)}}{J\sum\limits_{i=1}^{NP} x_i^2} + \frac{\sum\limits_{i=1}^{NP} x_i^2}{J\sum\limits_{i=1}^{NP} x_i^2}.\end{aligned} \tag{5.17}$$

4. Schritt: Zusammenfassung Schritt 3 in Vektorschreibweise

$$\widehat{\boldsymbol{\beta}}^{(b)} = \frac{1}{\|\boldsymbol{x}\|^2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y}^{(b)} + \mathbf{1}\left(\frac{\|\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y}^{(b)}\|_1}{J\|\boldsymbol{x}\|^2} + \frac{1}{J}\right). \tag{5.18}$$

Das Vorhersagemodell für den zu schätzenden Monat zum Zeitpunkt t_{fut} lautet damit

$$\boldsymbol{y} = x^{(t_{\text{fut}})}\left(\frac{1}{B}\sum_{b=1}^{B}\boldsymbol{\beta}^{(b)}\right)$$

mit

$$\widehat{\boldsymbol{\beta}}^{(b)} = \frac{1}{\|\boldsymbol{x}\|^2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y}^{(b)} + \mathbf{1}\left(\frac{\|\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y}^{(b)}\|_1}{J\|\boldsymbol{x}\|^2} + \frac{1}{J}\right). \tag{5.19}$$

5.3.5 Vor- und Nachteile

Auf den Hauptvorteil des Verfahrens wurde Eingangs des Abschnittes bereits hingewiesen. Statt weniger alter Referenzmonate existieren jetzt mehrere 100 Datensätze, die als Referenzmonate dienen. Dies ist ein explizites Einsatzgebiet für den Bootstrap. Für weitere Informationen zum Bootstrap und seine Anwendungsmöglichkeiten sei auf [EFR93] hingewiesen.
Durch die zufällige Verteilung der Passagiere einer Route ist zu erwarten, dass Routen mit wenigen Passagieren auch in den Datensätzen mit wenigen bis gar keinen Passagieren vertreten sind. Dies kommt zumindest häufiger vor, als dass eine Route mit wenigen Passagieren in einem Datensatz mit vielen Passagieren vertreten ist. Umgekehrt gilt derselbe Fall für Routen mit vielen Passagieren. Tendenzen in den Passagierzahlen werden damit gefestigt. Die Stabilität gegenüber Ausreißern wird erhöht.
Das Problem von Modell 2 sollte damit behoben sein. Es können nun ausschließlich aktuelle Daten verwendet werden.
Ein weiterer Vorteil dieses Verfahrens ist die bereits in Modell 2 angesprochene Geschwindigkeit die das lineare Modell mit sich bringt. Es gibt eine explizite Formel für $\widehat{\boldsymbol{\beta}}$, sodass kein Zeitaufwand für Näherungslösungen entsteht. In Kombination mit den wenigen Daten, die wegen der Anwendung des Perzentils übrig bleiben, kann B im oberen dreistelligen Bereich liegen, was die Stabilität massiv erhöht.
Die zu Modell 4 führende Erweiterbarkeit wurde bereits in Modell 2 dargelegt.

Probleme können in Bezug auf die Anzahl der Datensätze auftreten. Die Laufzeit steigt zwar lediglich linear, ungefähr in demselben Maße wie bei größerem B, allerdings werden nur sehr niedrige vierstellige Bereiche empfohlen. Es wird problematisch, wenn beide Werte B und N im oberen dreistelligen Bereich liegen. Auch eine größere Anzahl P an Referenzmonaten ist der Geschwindigkeit nicht zuträglich, wohl aber der Genauigkeit. Größer angelegte Tests sind nötig, um eine optimale Kombination aus B, N und

P zu bestimmen.
Des Weiteren sei erneut auf die Problematik der Anwendung des linearen Modells bei höherdimensionalen Problemen hingewiesen.

5.4 Modell 4: Einfache logistische Regression mit multiplen Input

Die Logistische Regression ist ein Modell aus dem Bereich der **Regressionsanalysen**. Dies sind Analyseverfahren der Statistik zur Modellierung der Beziehungen und Auswirkungen von einer oder mehrerer unabhängiger Variablen (den **Inputparametern**) und der oder den davon abhängigen Variablen (den **Outputparametern**). Im vorliegenden Fall wird das Modell als Prognoseverfahren verwendet. Andere Einsatzgebiete wie die Trennung von Funktion und Fehler (**Weißes Rauschen**) oder die quantitative Beschreibung derselben sind möglich. Ein anderer Name für die logistische Regression lautet auch **Logit-Modell**, wie im weiteren Verlauf der Modellbeschreibung noch erläutert wird. Im Gegensatz zu allen vorherigen Modellen ist nun die Möglichkeit gegeben, verschiedene Einflüsse zu verarbeiten. Jede Route besteht aus einzigartigen Eigenschaften, wie der Fluglänge, den verwendeten Flugzeugen, der Anzahl der Umstiege etc.. Diese Kriterien, die die Wahl eines einzelnen Passagiers für oder gegen eine Route in unbekanntem Maße beeinflussen, können nun allesamt in die Entscheidungsfindung einfließen. Die relative Häufigkeit der Passagiere einer Route wird damit nicht mehr direkt geschätzt, sondern die Wahrscheinlichkeit, dass eine Route überhaupt geflogen wird.

Das Modell wurde 1974 vom amerikanischen Ökonometriker Daniel McFadden und vom amerikanischen Ökonomen James Heckmann erdacht und weiterentwickelt. Es ist das wichtigste Verfahren für kategorielle abhängige Variablen. Ein eindrucksvolles Beispiel für seine Bedeutung ist die Tatsache, dass McFadden und Heckmann im Jahre 2000 für seine Entwicklung den Nobelpreis für Wirtschaftswissenschaften erhielten.
Das erste große Anwendungsgebiet lag im Bereich der Biomedizin. In den letzten 20 Jahren wurde die logistische Regression auch verstärkt in sozialwissenschaftlicher Forschung und Marketing eingesetzt. Ebenso bildet sie ein beliebtes Werkzeug in Businessanwendungen wie der Kreditbewertung. Mehr dazu siehe [AGR02].

Die einfache logistische Regression bildet die Vorstufe zu Modell 5 in Abschnitt 5.5, dem **Conditional Logit**, dessen Voraussetzungen und Einsatzgebiet besser mit dem vorgegebenen Problem harmoniert. Modell 4 wird aus zwei Gründen behandelt und implementiert. Zuallererst basiert es auf demselben theoretischen Fundament wie Modell 5 und eignet sich daher gut zur Vermittlung der Basiskenntnisse, vor allem da es das grundlegende Modell ist. Zum anderen sollte es ein Modell geben, mit dem die Güte des angepassteren Modells 5 verglichen werden kann.

5.4.1 Einführung

Wie im letzten Abschnitt bereits erwähnt wurde, besteht nun die Möglichkeit der Eingabe mehrerer Auswahlkriterien für eine Route, mit welcher aber nicht die relative Häufigkeit bestimmt wird, sondern die Wahrscheinlichkeit, dass sich ein einzelner Passagier überhaupt für diese Route entscheidet. Wohlgemerkt: *Ein einzelner Passagier*. In absoluten Zahlen lässt sich dann die Anzahl der Menschen bestimmen, die eine bestimmte Route fliegen oder nicht fliegen würden. Durch den Vergleich dieser Zahlen lässt sich im Endeffekt wieder die relative Häufigkeit der Passagiere einer Route bestimmen. Es ist wichtig, sich diesen Zusammenhang stets vor Augen zu halten, wenn im Text die Begriffe Wahrscheinlichkeit und relative Häufigkeit auftreten.
Konkret sieht die Idee zu dem Modell das folgende Anwendungsszenario vor: Aus den DPD, DD, Schedule- und den weiteren Daten aus Abschnitt 3 werden die Daten in Abschnitt 3.2.2 gewonnen. Die DPD dienen dabei als Outputparameter und die restlichen 20 Daten als Inputparameter. 20 und nicht 23, da die Werte x_{a1}, x_{a2} und x_{a3} von ihrer Erzeugung her mit dem Output gekoppelt sind und somit nicht für den Input verwendet werden können. Für jeden einzelnen Parameter wird über ein Anpassungsverfahren der Wert für einen expliziten Funktionsparameter bestimmt, der im Funktionsvektor $\boldsymbol{\beta}$ zusammengefasst ist. Dieser ist Hauptbestandteil einer Schätzfunktion π, deren Output den oben genannten Wahrscheinlichkeiten entspricht. Das Anpassungsverfahren versucht, diesen geschätzten Output an den tatsächlichen Output mittels Veränderung von $\boldsymbol{\beta}$ möglichst exakt anzupassen.
Nach der Beendigung der Anpassung liegt eine arbeitsfähige Schätzfunktion π vor. Für jede Route des zu schätzenden Monats sind nun die Inputparameter gemäß Abschnitt 3.2.2 zu bestimmen und einzugeben (ohne x_{a1}, x_{a2} und x_{a3}). Der gewonnene Outputwert entspricht der Wahrscheinlichkeit, dass sich ein einzelner Passagier für diese Route entscheidet. Die Berechnung der Wahrscheinlichkeit erfolgt unabhängig von den anderen Routen. Sie sind in der Formel nicht enthalten. Damit ergeben alle Wahrscheinlichkeiten in der Summe einen Wert ungleich eins. Durch eine Normierung der Wahrscheinlichkeiten aller Routen eines Weges zu eins wandelt sich $\pi(x_i)$ zur eigentlich gesuchten relativen Häufigkeiten der DPD.
Herleitung und Argumentation sind an [AGR02] angelehnt.

Bemerkung 5.4.1
Die Begriffe „Variablen“ und „Parameter“ werden synonym verwendet. Zur Vermeidung von Inkonsistenz in der Namensgebung und Irritationen bei Vergleichen mit Fachliteratur wird im Nachfolgenden daher lediglich von „Parametern“ gesprochen.

Die weitere Arbeit führt zuerst die logistische Regression mit einem einzigen Inputparameter x_i ein. Die Erweiterung auf einen multiplen Inputvektor $\boldsymbol{x}_i$ ist leicht zu bewerkstelligen und wird im Text ausdrücklich erwähnt, sodass der Zeitpunkt erkennbar ist, ab dem von höheren Dimensionen die Rede ist (konkret ist dies Abschnitt 5.4.6).

Genauer gesagt dient das Modell als Regressionsanalyse zur Modellierung der Verteilung diskreter abhängiger Outputparameter. Liegen davon mehrere vor, so wird von **Multinomialer Logistischer Regression** gesprochen. Sie dient allgemein zur Schätzung von Gruppenzugehörigkeiten. Im vorliegenden Fall handelt es sich um eine **binäre logistische Regression** für relle abhängige Inputparameter. Binär bedeutet, dass eine Beobachtung i aus einem Input-Output-Paar $(x_i, y_i)_{i=1,\dots,N}$ mit $y_i \in \{0,1\}$ besteht.

Definition 5.4.1 (Beobachtung)
*Eine **Beobachtung** i besteht aus einem **Input-Output-Paar** $(x_i, y_i)_{i=1,\dots,N}$. N steht für die Mächtigkeit der Gesamtmenge an Beobachtungen. y_i heißt **Regressand** und bildet den binären abhängigen Outputparameter für den Inputparameter, den bekannten und festen Kovariablenvektor x_i.*
Beschreiben W_x den Wertebereich der x_i und D_y den Definitionsbereich der y_i, so wird die Beziehung des Input-Outputpaares über die Funktion

$$\pi(x_i) = y_i \text{ mit } \pi : W_x \to D_y = \{0,1\}$$

dargestellt. Im späteren Verlauf der Arbeit wird D_y über die natürlichen Zahlen definiert.

Achtung: Im Modelltraining besteht eine Beobachtung aus einem Input-Output-Paar mit bekanntem y_i. Für die spätere Schätzung ist y_i unbekannt und über das Modell zu bestimmen.

Bemerkung 5.4.2
Bei der Verwendung des Modells muss zwischen Theorie und Anwendung unterschieden werden. Die Werte der y_i sind bekannt und $\in \{0,1\}$. Die Werte der $\pi(x_i)$ sind Schätzungen und $\in [0,1]$ und müssten theoretisch mittels eines Schwellwertes zu 0 und 1 zugeordnet werden.
Aufgrund der Wichtigkeit der Aussage, sei der Absatz aus dem Einführungstext noch einmal wiederholt, um Irritationen zu vermeiden. $\pi(x_i)$ entspricht der Wahrscheinlichkeit, dass eine Route geflogen wird oder nicht, unabhängig von der Anzahl der anderen möglichen Routen. Diese Aussage würde sich mittels eines Schwellwertes und der Zuordnung zu 0 oder 1 konkret in die Aussagen „Route wird geflogen“, „Route wird nicht geflogen“ kategorisieren lassen. Durch eine Normierung der Wahrscheinlichkeiten aller Routen eines Weges zu 1 wandelt sich $\pi(x_i)$ zur eigentlich gesuchten relativen Häufigkeiten der DPD.

5.4.2 Das Logit-Modell

Laut [AGR02] sind nichtlineare Beziehungen zwischen $\pi(x)$ und x häufig monoton. Mit dem Wachstum von x folgt ein kontinuierliches Wachstum oder Fall von $\pi(x)$. Das Ergebnis ist eine S-förmige Kurve, deren steigender Verlauf untenstehendem Beispiel 5.4.1 entsprechen würde.

Eine Funktion die ein derartiges Aussehen besitzt und, anders als viele lineare Modelle, eine garantierte Abbildung in das Intervall $[0,1]$ liefert, ist die für das Modell namensgebende **Logistische Funktion** $f(x) = \frac{1}{1+e^{-x}}$

In der Arbeit wird eine leicht abgeänderte Darstellung genutzt $f(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$.

Beispiel 5.4.1
Laut dem linearen Modell ist der Einfluss eines monatlichen Mehrwertes an Vermögen von $50 €$ bei jeglichem monatlichen Verdienst x gleich. In der Praxis ist es aber ein Unterschied, ob ein Geringverdiener mit einem Monatseinkommen von $x = 800 € 50 €$ und damit 6.25% mehr erhält, als ein Gutverdiener mit $x = 5\,000 €$ und damit 1% mehr im Portemonnaie hat. Ein solcher nichtlinearer Verlauf tritt in der Praxis häufiger auf, als ein steifer linearer Verlauf. Vor allem, wenn die Anzahl der Inputparameter groß ist und zusätzlich noch korreliert.

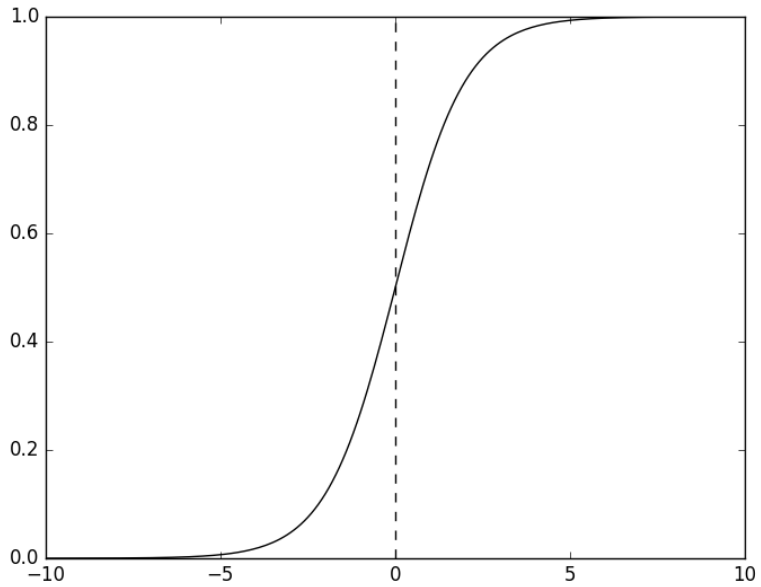


Abbildung 5.1: Die Logistische Funktion

Es werden zudem weitere Parameter benötigt. Diese sind

- α, für die Darstellung des **Weißen Rauschens**, siehe Kapitel 4;
- **Anpassungskoeffizient** β, für die Beschreibung der Stärke des Einflusses von x.

Durch das Ersetzen von x mittels der Linearkombination von α und β entsteht das **Logit Modell**, das Modell der einfachen logistischen Regression:

Definition 5.4.2 (Logit Modell)
Sei X eine erklärende Zufallsvariable zu einer zweiteiligen Beobachtung aus einer Menge von N Beobachtungen, welche aus einem Vektor mehrerer Variablen $x_1, ..., x_p$ bestehen kann. Aktuell sei nur von einer Variablen x ausgegangen. Sei Y die Responsevariable zu X, welche den zweiten Teil der Beobachtung stellt. Für eine beliebige Beobachtung ordnet $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$ der Variablen X die Eintrittswahrscheinlichkeit des Wertes 1 der Responsevariablen zu. Besitzt die Responsefunktion π das Aussehen

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}, \tag{5.20}$$

*so wird vom **Modell der einfachen logistischen Regression** oder auch dem **Logit Modell** gesprochen.*

- *Für $\beta < 0$ und $x \to \infty$ gilt $\pi(x) \to 0$.*
- *Für $\beta > 0$ und $x \to \infty$ gilt $\pi(x) \to 1$.*

Damit kann der komplette Wertebereich der reellen Zahlen abgedeckt werden.

*Die Wahrscheinlichkeit für $Y = 1$ wird dabei nicht aus der erklärenden Variable modelliert, sondern mittels des **Logit**. Dabei handelt es sich um die logarithmierte Wahrscheinlichkeit für $Y = 1$*

$$\alpha + \beta x = \text{Logit}\,(Y = 1|X = x) = \ln \frac{\pi(x)}{1 - \pi(x)}. \tag{5.21}$$

5.4.3 Verbindung zum linearen Modell

Dieser Abschnitt soll die Verbindung des Logit-Modells zu den allgemeinen linearen Modellen aufzeigen, deren Weiterentwicklung der Logit darstellt, welcher die Verbindung zum Logit-Modell liefert. Dazu müssen die sogenannten **Odds** betrachtet werden.

Definition 5.4.3 (Odds)
Odds („Chancen“, „Quoten“) sind eine von der Statistik gegebene Möglichkeit zur Darstellung von Wahrscheinlichkeiten des Eintretens eines Ereignisses A. Gebildet werden Odds als Quotient der Eintrittswahrscheinlichkeit P von A und der Gegenwahrscheinlichkeit des Nichteintretens

$$R(A) = \frac{P(A)}{1 - P(A)}. \tag{5.22}$$

Werte > 1 bedeuten größere Wahrscheinlichkeit für Eintreten von Ereignis A, wohingegen Werte < 1 für eine erhöhte Wahrscheinlichkeit des Nicheintretens von Ereignis A stehen.

Die Kenntnis über die Odds liefert die Eintrittswahrscheinlichkeit von A

$$P(A) = \frac{R(A)}{1+R(A)}. \tag{5.23}$$

Modelle der Statistik können demzufolge auch allein mittels $R(A)$ bearbeitet werden.

Logarithmierte Odds werden kurz auch **Log-Odds** genannt. Die Log-Odds des Logit-Modells liefern folgende interessante lineare Beziehung

$$\text{logit}\,(\pi(x)) = \log \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x. \tag{5.24}$$

Die Umwandlung ist die sogenannte **Log-Odds-Transformation**. Diese spezielle Darstellung wird als **Logit** bezeichnet.

$\pi(x)$ ist der natürliche Parameter der Binomialverteilung. Damit bildet die Log-Odds-Transformation ihren kanonischen Link. Während $\pi(x)$ Werte des Intervalls $[0,1]$ annimmt, bilden die reellen Zahlen den Wertebereich des Logits, ebenso wie bei linearen Vorhersagemodellen wie $\alpha + \beta x$, welche die systematische Komponente allgemeiner linearer Modelle bilden. Damit hat das Modell der logistischen Regression nicht dieselben strukturellen Probleme wie das lineare Wahrscheinlichkeitsmodell [AGR02].

5.4.4 Der Einfluss von β und Eigenschaften von π

Dieser Abschnitt soll den Einfluss von β auf den Verlauf der Funktion des logistischen Modells π mit einem eindimensionalen Inputparameter x, sowie einige geometrische Eigenschaften von π behandeln.

- Das Vorzeichen von β entscheidet, ob $\pi(x)$ bei wachsendem x steigt oder fällt.
- Der Grad des Anstiegs wird durch $|\beta|$ bestimmt.
  - Für $\beta \to \infty$ nähert sich $\pi(x)$ der Heaviside-Funktion an.
  - Für $\beta \to 0$ nähert sich $\pi(x)$ einer horizontalen Linie auf Höhe $\frac{e^\alpha}{1+e^\alpha}$ an. Gleichheit bedeutet, dass $\pi(x)$ unabhängig von x ist und das Input-Output-Paar damit in keinerlei Relation steht.
- Sei $\beta > 0$ und $x \in \mathbb{R}$ so besitzt $\pi(x)$ die Form der logistischen Funktion. Diese Funktion ist punktsymmetrisch im Punkt $(x, \pi(x))$ für x mit $\pi(x) = 0.5$. Die Annäherung an 1 erfolgt in derselben Rate wie bei Annäherung an 0 (ähnlich wie bei der logistischen Funktion in Abbildung 5.1)
- Log-Odds sind Exponentialfunktionen von x, sichtbar durch Exponenzierung der Logitgleichung. Es lässt sich folgende Aussage für die Größe von β treffen: Die Odds wachsen multiplikativ um e^β, wenn x um die Einheit 1 wächst.

- Eine andere Möglichkeit der Anstiegsdarstellung wurde durch Berkson 1951 aufgezeigt [BER51]. Sie erfolgt durch die typische Betrachtung des Tangentenanstiegs von π an einer Stelle x durch Ermittlung der Ableitung. $\frac{\partial \pi(x)}{\partial x}$ liefert nach Vereinfachung eine Änderungsrate von $\beta \pi(x)\,(1-\pi(x))$ (siehe Abbildung 5.2).
- α als Manifestierung des weißen Rauschens ist beim einfachen logistischen Modell nicht weiter von Belang. Lediglich bei der Zentrierung von π an der Stelle $x = 0$ (mittels Substitution von x durch $(x-\bar{x})$), wird α zum Logit an dieser Stelle, sodass $\pi(\bar{x}) = \frac{e^{\alpha}}{1+e^{\alpha}}$. Dies ist wichtig für spezielle Formen des logistischen Modells, bei denen quadratische Terme oder Interaktionsterme zur Reduzierung von Korrelationen der Modellparameter enthalten sind. Dieses Vorgehen wird gern bei gewöhnlicher Regression oder in anderen komplexen Modellen genutzt.

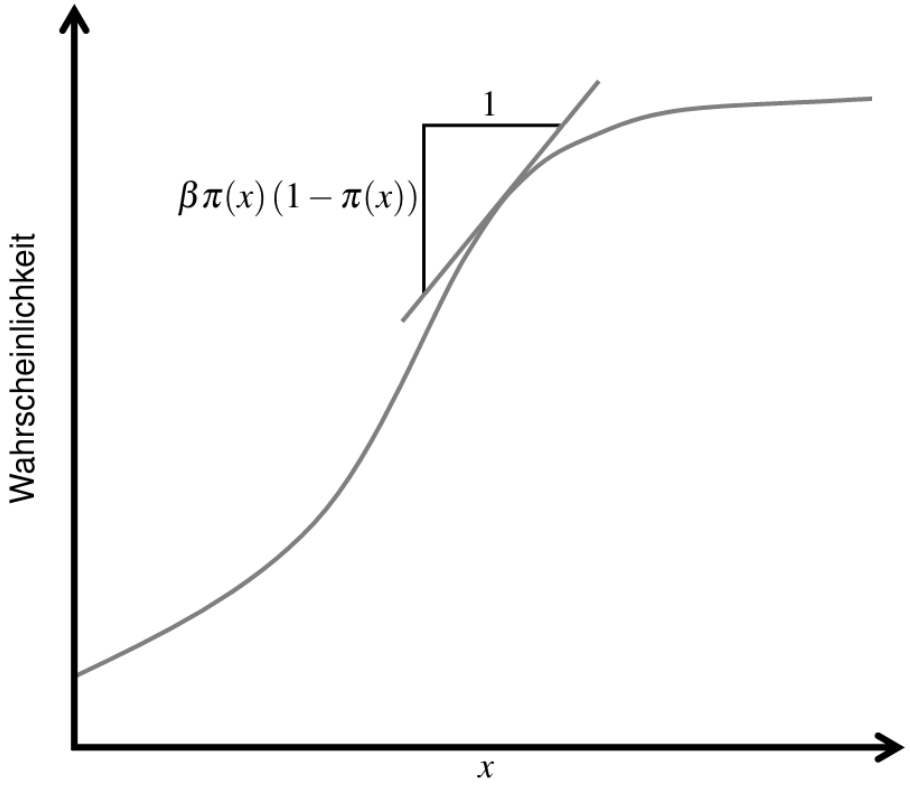


Abbildung 5.2: Lineare Approximation der logistischen Regressionskurve

5.4.5 Umwandlung zum logistischen Regressionsmodell mit multiplen Input

Das bisherige Modell der einfachen logistischen Regression betrachtete lediglich einen Einflussparameter. Ab sofort wird das Modell der logistischen Regression mit multiplem Inputvektor $\boldsymbol{x}$ und damit mit einem Inputvektor $\boldsymbol{\beta}$ arbeiten. Nun existieren mehrere Inputparameter, also verschiedene Einflüsse, die in das Modell, die Formel und das Gesamtkonzept integriert werden müssen. Die Art und Beschaffenheit der Inputparameter ist in

Abschnitt 3.2.2 angegeben. Die Werte x_{a1}, x_{a2} und x_{a3} werden dabei nicht aufgenommen, da sie für den zu schätzenden Monat aller Voraussicht nach nicht vorhanden sind und damit nicht über einen Koeffizienten β_k bewertet werden können. Eine Aufnahme dieser Daten in den Input des Trainingsprozesses würde damit nur zur unnötigen Verschiebungen der Prioritäten in der Anpassung führen.
Wie im vorherigen Modell wird für *jeden* Weg *ein* spezifisches $\boldsymbol{\beta}$ bestimmt. Es reicht somit, die Modellschätzung für *einen* Weg zu betrachten. Zur besseren Lesbarkeit entfällt ein entsprechender Index in den verwendeten Variablen.
Betrachtet sei die Beobachtung i. Ihr Inputvektor $\boldsymbol{x}$ zeichnet sich dadurch aus, dass er Eigenschaften *einer* bestimmten Route zu *einem* bestimmten Zeitpunkt enthält. Zu jedem Inputparameter x_{ik} existiert ein Koeffizent β_k, der in bereits ausgeführter Art und Weise die Stärke des Einflusses von x_{ik} auf $\pi(\boldsymbol{x}_i)$ angibt (siehe dazu Abschnitt 5.4.4). Die Gesamtheit der β_k wird als Koeffizientenvektor $\boldsymbol{\beta}$ beschrieben. In π bilden die β_k und x_{ik} eine Linearkombination, die sich auch als Skalarprodukt der beiden Vektoren darstellen lässt. α kann in diesem Skalarprodukt eingefügt werden, indem $\boldsymbol{x}$ um das Element 1 an der ersten Stelle erweitert wird; α ist in $\boldsymbol{\beta}$ an derselben Stelle einzubetten, an der die 1 in $\boldsymbol{x}$ integriert wurde. Der Wert von α wird automatisch bei der Bestimmung konkreter Werte von $\boldsymbol{\beta}$ ermittelt.

Bei Einführung dieser Erweiterungen ergibt sich das Modell dieses Abschnitts, das **Modell der logistischen Regression mit multiplen Input**

$$\pi(\boldsymbol{x}_i) = \frac{e^{\sum_{k=1}^{p} \beta_k x_{ik}}}{1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}}} = \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}. \tag{5.25}$$

Zur Erinnerung an Abschnitt 5.4.1 seien noch einmal die wichtigsten Punkte zu $\pi(x_i)$ aufgeführt:

1. $\pi(x_i)$ gibt im Sinne der Anwendung für das vorliegende Problem die Wahrscheinlichkeit für einen einzelnen Passagier an, ob er mit der eingegebenen Route x_i fliegt oder nicht. Diese Wahrscheinlichkeit gilt für alle Passagiere, die diese Route fliegen wollen.

2. Für nachfolgend erklärte Anpassung der β_k werden alle verfügbaren und sinnvollen Routen zu allen verfügbaren und sinnvollen Zeiten *eines bestimmten Weges* zum Training benutzt. Es ergibt sich folgende Änderung in der Idee der Modellanwendung: Nach dem Modelltraining wird jede Route eines bestimmten Weges zum Vorhersagezeitpunkt t eingegeben und die entsprechenden Wahrscheinlichkeiten ermittelt. Dies entspricht nicht ganz der Verteilung der Passagiere eines Weges auf die jeweilige Route in Form der relativen Häufigkeiten, da die Berechnung unabhängig von konkurrierenden Routen erfolgt, womit sich die ermittelten Wahrscheinlichkeiten nicht zu eins addieren. Als Lösung für dieses Modell wer-

den die Routenwahrscheinlichkeiten der Routen eines Weges genormt, sodass sich die Summe der genormten Wahrscheinlichkeiten zu eins addiert. Das Ergebnis ist die geforderte relative Häufigkeit. Dieses Vorgehen ist möglich, da der Output im Intervall $[0,1]$ liegt und somit keine Verfälschungen durch negative Werte auftreten kann.

5.4.6 Training des Modells - Anpassung von $\boldsymbol{\beta}$

In den vorherigen Abschnitten wurde die Herleitung des Modells behandelt. Der vorliegende Abschnitt ist der Anpassung von $\boldsymbol{\beta}$ gewidmet. Sobald $\widehat{\boldsymbol{\beta}}$ ermittelt ist, sind für den Vorhersagezeitpunkt t pro Weg alle Routen in das Modell einzugeben (dies erfolgt mittels ihrer Routeneigenschaften), womit sich die Verteilung der Passagiere auf die Routen bestimmen lässt.
Die Anpassung von $\boldsymbol{\beta}$ besteht aus zwei Schritten. Der zweite Schritt benötigt einen Extraschritt, welcher wiederum einen Extraschritt braucht:

1. Zunächst wird eine Zielfunktion, die **Log-Likelihood-Funktion**, bestimmt.

2. Die Schätzung $\widehat{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$ erfolgt mittels **Maximum-Likelihood-Methode**. Die Log-Likelihood-Funktion wird mittels eines Startwertes ausgewertet.

    - Das Ergebnis der Bearbeitung ist ein nichtlineares Gleichungssystem welches per iterativer Optimierung über das **Newton-Raphson-Verfahren** gelöst wird.

        – In jedem Iterationsschritt ist die Inverse einer Hessematrix zu berechnen, deren Existenz nicht vorauszusetzen ist. Stattdessen wird ein lineares Gleichungssystem per **Newton-konjugiertem Gradientenabstieg** gelöst.

Die Schritte 1-3 sind aus [AGR02] und der vierte Schritt in [HAN09] entnommen.

Ermittlung der Likelihood-Funktion

Zur Erstellung der Likelihood-Funktion liefert [AGR02] in Abschnitt 5.5.1 folgende Möglichkeit zu einer vereinfachenden Annahme: Bei multiplem Auftreten von Beobachtungen mit binärem Output und identischem Input $\boldsymbol{x}_i$ ist es hinreichend, die Anzahl n_i dieser Beobachtungen und die Anzahl der Erfolge zu ermitteln. Somit ist nur noch eine Beobachtung mit Input $\boldsymbol{x}_i$ und und Output y_i vorhanden, wobei y_i nun auf die Anzahl der Erfolge verweist. Damit werden die bernoulliverteilten Einzelerfolge zu unabhängig binomialverteilten Outputs $\{y_1,...,y_N\}$ mit $E(y_i) = n_i\pi(x)_i$. Im Rahmen der vereinfachten

Annahme steht n_i für die Anzahl der Wegpassagiere, welche zum Zeitpunkt der Beobachtung die Wahl hatten, sich für die zum Input $\boldsymbol{x}_i$ gehörende Route zu entscheiden.

Das Produkt von N Binomialfunktionen ist damit proportional zur vereinigten Likelihood-Funktion der Wahrscheinlichkeiten

$$\begin{aligned}\text{Likelihood-Funktion} &= \prod_{i=1}^{N} \pi(\boldsymbol{x}_i)^{y_i} \left(1-\pi(\boldsymbol{x}_i)\right)^{n_i-y_i} \\ &= \left(\prod_{i=1}^{N} e^{\log\left(\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right)^{y_i}\right)} \right) \left(\prod_{i=1}^{N} \left(1-\pi(\boldsymbol{x}_i)\right)^{n_i} \right) \\ &= \left(e^{\sum_{i=1}^{N} y_i \log\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right)} \right) \left(\prod_{i=1}^{N} \left(1-\pi(\boldsymbol{x}_i)\right)^{n_i} \right). \end{aligned} \tag{5.26}$$

Variablen

Mit der Einführung eines neuen Verständnisses der Outputvariable einer Beobachtung geht auch ein neues Verständnis der bisher verwendeten Variablen einher. Diese wie auch neue sind in Tabelle 5.4 aufgeführt, welche bereits ab Abschnitt 5.4.6 gelten.
Wie im vorherigen Modell wird für *jeden* Weg *ein* spezifisches $\boldsymbol{\beta}$ bestimmt. Es reicht somit, die Modellschätzung für *einen* Weg zu betrachten. Zur besseren Lesbarkeit entfällt damit ein entsprechender Index.
Es sei darauf hingewiesen, dass für eine Route mehrere Beobachtungen existieren können, wenn die Route über mehrere Zeitpunkte hinweg untersucht/beobachtet wird.

Variable	Beschreibung
t	Iterationsschritt des Newton-Raphson-Verfahrens, er betrifft $\boldsymbol{\beta}$ und damit auch π, u und H; wird als obere Indexmarkierung angegeben
N	Anzahl der Beobachtungen über mehrere eingegebene Jahre $=$ Anzahl der Routen
$i = 1, ..., N$	Index für eine spezielle Beobachtung
p	Anzahl der Inputparameter
$k = 1, ..., p$	Index für einen speziellen Inputparameter
$\boldsymbol{x}_i$	$1 \times p$ - Inputvektor, welcher die Eigenschaften der Route der Beobachtung i enthält; bestehen aus den Variablen aus Abschnitt 3.2.2
x_{ik}	Eine einzelne Eigenschaft von $\boldsymbol{x}_i$ (ohne x_{a1}, x_{a2} und x_{a3})
$\boldsymbol{X}$	$N \times p$ - Inputmatrix, die alle Inputvektoren aller Beobachtungen enthält
y_i	Output einer Route der Beobachtung i; besteht aus der Anzahl der Passagiere, welche die Route zum Zeitpunkt der Beobachtung in Anspruch genommen haben
$\boldsymbol{y}$	$1 \times N$ - Outputvektor der alle Outputs eines Weges enthält
$\boldsymbol{\beta}$	$1 \times p$ - Vektor der Modellparameter
$\pi(x_i)$	Wahrscheinlichkeit einer Route für ihren Flug, geschätzter Output
$\boldsymbol{\pi}$	$1 \times N$ - Vektor, der den geschätzten Output der Routen angibt
n_i	Anzahl der Passagiere des Weges einer Beobachtung i zum Zeitpunkt dieser Beobachtung; Es gilt $n_j = n_h$ für Beobachtungen h und j aus demselben Jahr und $n_j \neq n_h$ für Beobachtungen h und j aus verschiedenen Jahren, bei denen die Wegpassagierzahlen unterschiedlich waren
$\boldsymbol{n}$	$1 \times N$ - Vektor, der die Gesamtanzahl der Wegpassagierealler Beobachtungen in ihrer Auftrittsreihenfolge in $\boldsymbol{X}$ enthält
$L(\boldsymbol{\beta})$	Logarithmische Zielfunktion, die nach $\boldsymbol{\beta}$ optimiert werden soll
u_a	Erste Ableitung der Zielfunktion nach Parameter β_a
$\boldsymbol{u}$	$1 \times p$ - Vektor der ersten Ableitung der Zielfunktion nach allen Parametern
h_{ab}	Zweite Ableitung der Zielfunktion nach den Parametern β_a und β_b
$\boldsymbol{H}$	$p \times p$ - Hessematrix der zweiten Ableitung der Zielfunktion nach allen Parametern
diag $[\,]$	$N \times N$ - Diagonalmatrix mit den Eintragungen an allen Stellen (i, i) wie in der Klammer beschrieben
$\circ$	Vektormultiplikationszeichen, welches für eine elementweise Multiplikation steht

Tabelle 5.4: Variablen dieses Abschnittes

Ermittlung der Log-Likelihood-Funktion

Für das weitere Vorgehen sind folgende zwei leicht nachvollziehbare Aussagen von Nöten

$$e^{\sum_{i=1}^{N} y_i \left(\sum_{k=1}^{p} \beta_k x_{ik} \right)} = e^{\sum_{k=1}^{p} \left(\sum_{i=1}^{N} y_i x_{ik} \right) \beta_k} \tag{5.27}$$

und

$$1 - \pi(\boldsymbol{x}_i) = \left(1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}} \right)^{-1}. \tag{5.28}$$

Durch Logarithmierung der Likelihood-Funktion (5.26) und Ersetzen der entsprechenden Stellen durch die Terme (5.27) und (5.28) entsteht die **Log-Likelihood-Funktion**

$$L(\boldsymbol{\beta}) = \sum_{k=1}^{p} \left(\sum_{i=1}^{N} y_i x_{ik} \right) \beta_k - \sum_{i=1}^{N} n_i \log \left(1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}} \right). \tag{5.29}$$

Die Log-Likelihood-Funktion ist die logarithmierte Form der Likelihood-Funktion. Die **Maximum-Likelihood-Methode** schätzt die unbekannten Parameter $\boldsymbol{\beta}$ als die Parameter, welche die Likelihood-Funktion maximieren.

Parameterschätzung mittels Maximum-Likelihood-Methode

Die Grundidee für die Maximum-Likelihood-Methode ist einfach: Ein gesuchter Parameter wird geschätzt, indem eine Wahrscheinlichkeits- oder andere Funktion über diesem Parameter nach dem Maximum optimiert wird.
Sie ist ein 1922 vom britischen Genetiker, Evolutionstheoretiker und Statistiker Sir Ronald Aylmer Fisher beschriebenes parametrisches Schätzverfahren der Statistik. Es ist derjenige Parameter als Schätzung auszuwählen, bei dem gemäß seiner Verteilung die vernünftigste Realisierung der Beobachtungsdaten eintritt. Im vorliegenden Fall fällt die Wahl auf den einzigen unbekannten Parameter $\boldsymbol{\beta}$. Liegt eine von $\boldsymbol{\beta}$ abhängige Wahrscheinlichkeitsdichte mit

$$\pi : \Omega \to [0,1], \ \ \boldsymbol{x} \mapsto \pi(\boldsymbol{x} \mid \boldsymbol{\beta})$$

vor, so wird die Likelihood-Funktion zu einem beobachteten Ausgang $\boldsymbol{x}$ für verschiedene Parameterwerte bestimmt

$$L : \Theta \to [0,1], \ \ \boldsymbol{\beta} \mapsto \pi(\boldsymbol{x} \mid \boldsymbol{\beta}).$$

Hat $\boldsymbol{\beta}$ einen bestimmten Wert, so liefert die Likelihood-Funktion die Wahrscheinlichkeit der Beobachtung des Ereignisses $\boldsymbol{x}$. Das $\boldsymbol{\beta}$, welches die Likelihood-Funktion maximiert, heißt **Maximum-Likelihood-Schätzung**.
Die Maximum-Likelihood-Methode ist das wichtigste Verfahren zur Erzeugung von Schätz-

funktionen für die Parameter einer Verteilung.
Die Maximierungsaufgabe kann laut der Theorie von Ockhams Rasiermesser [OCC] mit Hilfe der einfachen Schulmathematik angegangen werden. Dies bedeutet die Ableitung der Log-Likelihood-Funktion nach dem gesuchten Parameter, im vorliegenden Fall nach $\boldsymbol{\beta}$. Da überwachtes Lernen möglich ist, sind die restlichen Variablen x und y bekannt. Es folgt das Nullsetzen des entstandenen Ableitungsterms, die Umstellung nach $\boldsymbol{\beta}$ und das Einsetzen in die zweite Ableitung zur Kontrolle des Extremums auf Maximum oder Minimum.
Die erste partielle Ableitung nach dem a-ten Element von $\boldsymbol{\beta}$ lautet

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_a} = u_a(\boldsymbol{\beta}) = \sum_{1=1}^{N} y_i x_{ia} - \sum_{i=1}^{N} n_i x_{ia} \frac{e^{\sum_{k=1}^{p} \beta_k x_{ik}}}{1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}}}. \tag{5.30}$$

In anderer Schreibweise

$$\boldsymbol{u} = (\boldsymbol{y} - \boldsymbol{n} \circ \boldsymbol{\pi})\boldsymbol{X}. \tag{5.31}$$

Die zweite partielle Ableitung nach dem a-ten und b-ten Element von $\boldsymbol{\beta}$ lautet

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_a \partial \beta_b} = h_{ab}(\boldsymbol{\beta}) = -\sum_{i=1}^{N} n_i x_{ia} x_{ib} \frac{e^{\sum_{k=1}^{p} \beta_k x_{ik}}}{\left(1 + e^{\sum_{k=1}^{p} \beta_k x_{ik}}\right)^2}. \tag{5.32}$$

Dies sind die Elemente der **Hessematrix**. In anderer Schreibweise

$$\boldsymbol{H} = -\boldsymbol{X}^{\mathsf{T}} \mathbf{diag}\,[n_i \pi(\boldsymbol{x}_i)(1 - \pi(\boldsymbol{x}_i))]\boldsymbol{X} \qquad i \in (1, ..., N). \tag{5.33}$$

Die Likelihood-Gleichung lautet

$$\boldsymbol{u}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{n} \circ \boldsymbol{\pi})\boldsymbol{X} = \mathbf{0}. \tag{5.34}$$

Damit gilt

$$\boldsymbol{y}\boldsymbol{X} = (\boldsymbol{n} \circ \boldsymbol{\pi})\boldsymbol{X}. \tag{5.35}$$

Da die linke Seite im Training bekannt ist, handelt es sich um eine nichtlineare Gleichung die einer iterativen Näherungslösung bedarf. Die linke Seite liefert nebenbei die Begründung dafür, dass $\boldsymbol{\pi}$ als Näherungslösung für $\boldsymbol{y}$ angesehen wird (genauer gesagt $\boldsymbol{n} \circ \boldsymbol{\pi}$).

Lösung der Likelihood-Gleichung mittels Newton-Raphson-Verfahren

Zur Lösung der nichtlinearen Gleichung (5.35) kann laut [AGR02] das Newton-Raphson-Verfahren verwendet werden. Dieses Verfahren ist benannt nach dem englischen Na-

turforscher Sir Isaac Newton (1736 „Method of Fluxions“, bereits 1671 geschrieben) und dem englischen Mathematiker Joseph Raphson (1690 „Analysis aequationum universalis“), welcher das Manuskript Newtons einsehen durfte. Siehe [DEU02], [DEU04], [ORT00] und [NEW]. Es ist geeignet für die Lösung der Likelihood-Gleichung, da es sich um eine iterative Methode zur numerischen Lösung nichtlinearer Gleichungen beziehungsweise Gleichungssysteme handelt, respektive zur näherungsweisen Ermittlung derer Nullstellen, welche gesucht sind. Aufgrund der verwendeten Binomialform und den Eigenschaften der darin verwendeten Exponentialfunktion existiert nur ein Extremum, das globale Maximum der Log-Likelihood-Funktion.

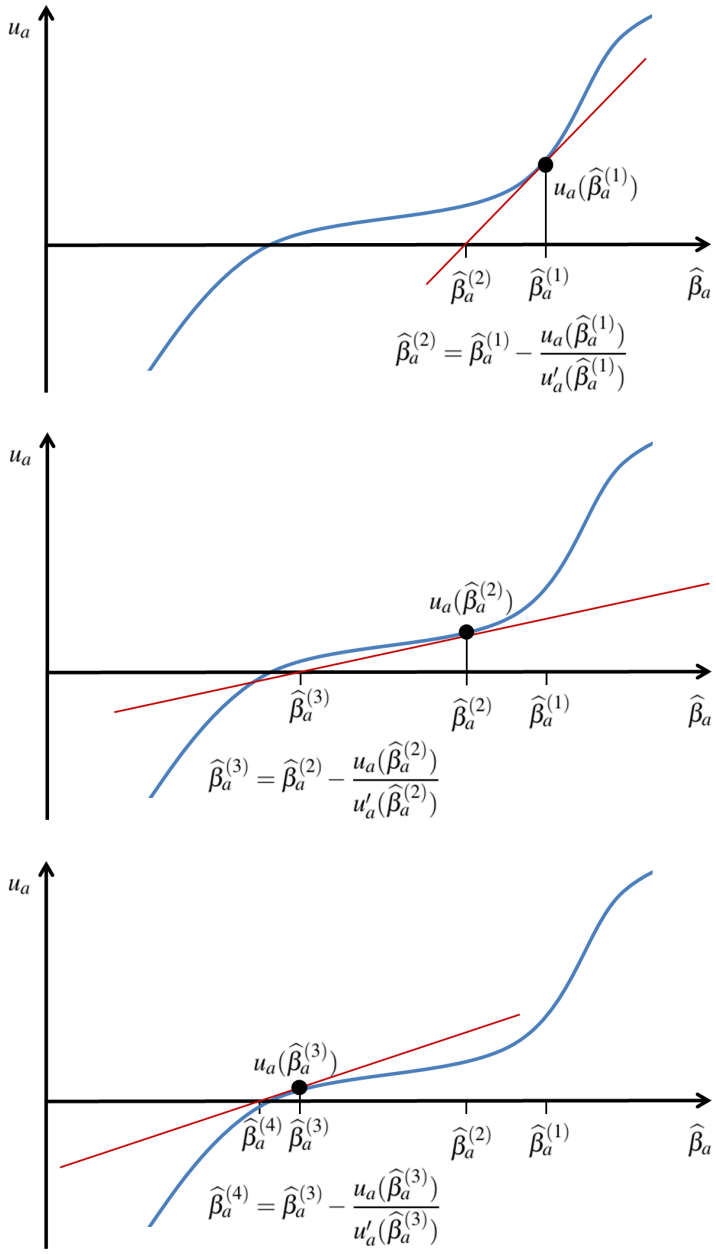


Abbildung 5.3: Darstellung der Funktionsweise des Newton-Raphson-Verfahren für eindimensionalen Input. Es sind drei Iterationschritte abgebildet.

Idee mit eindimensionalen Input

Für die allgemeine Lösungsidee ist die Vorbereitung bereits abgeschlossen. Sie wird im Folgenden für den eindimensionalen Input beschrieben, um Bilder zur Verdeutlichung liefern zu können und beschäftigt sich mit der Ableitung.

Die Idee lautet: Die stetig differenzierbare reelle Funktion $u_a : \mathbb{R} \to \mathbb{R}$ ist zu vereinfachen, indem an gewisser Stelle $\widehat{\beta}_a$ die Funktion u_a an der Stelle $\widehat{\beta}_a$ durch eine Tangente mit dem Anstieg $u'_a(\widehat{\beta}_a)$ ersetzt wird. Gesucht ist nun die Nullstelle, die in der Log-Likelihood-Funktion eine Extremstelle darstellt. Diese Nullstelle bildet den Ausgangspunkt für den nächsten Iterationsschritt. Das Vorgehen ist für die ersten drei Schritte in der Abbildung 5.3 skizziert.

Verfahren mit eindimensionalen Input

Im eindimensionalen Fall ist die Tangente mittels

$$\text{Tangente}(\beta_a) = u_a(\widehat{\beta}_a) + u'_a(\widehat{\beta}_a)(\beta_a - \widehat{\beta}_a) \tag{5.36}$$

beschrieben, wobei $u_a(\widehat{\beta}_a)$ die Verschiebung entlang der Ordinate und $-\widehat{\beta}_a$ die Verschiebung entlang der Abszisse darstellt.

Für die Iteration wird $\widehat{\beta}_a$ durch $\widehat{\beta}_a^{(t)}$ und β_a durch $\widehat{\beta}_a^{(t+1)}$ ersetzt. Damit ergibt sich eine Rekursionsvorschrift für eine unendliche Folge von Stellen $(\widehat{\beta}_a^{(t)})_{t\in\mathbb{N}}$

$$\widehat{\beta}_a^{(t+1)} = \widehat{\beta}_a^{(t)} - \frac{u_a(\widehat{\beta}_a^{(t)})}{u'_a(\widehat{\beta}_a^{(t)})}. \tag{5.37}$$

Dieses Vorgehen trägt den Namen **Newton-Iteration**.

Bemerkung 5.4.3

Als Spezialfall einer Fixpunktiteration ist $u_a(\xi) = 0$, da $\xi = \xi - \frac{u_a(\xi)}{u'_a(\xi)}$ falls $\lim_{t\to\infty} \widehat{\beta}_a^{(t)} = \xi$.

Als sogenanntes lokal konvergentes Verfahren ist die Konvergenz nur bei hinreichender Nähe des Startwertes an der Nullstelle gegeben. Im gegenteiligen Fall kann ein schlechter Startwert zur Divergenz oder Oszillation der Folge führen. Im günstigen Fall konvergiert das Verfahren quadratisch.

Falls es jedoch im Intervall $I =]a;b[$ genau eine Nullstelle gibt, in I durchweg $f' > 0$ sowie $f'' < 0$ gilt und der Startwert $x_0 \in I$ links von der Nullstelle $\xi \in I$ gewählt wird, dann konvergiert die Folge im Newton-Verfahren stets, und zwar streng monoton wachsend (wie bei Abbildung 5.3). Wird der Startwert rechtsseitig gewählt, so drehen sich die Größenrelationszeichen um.

Bemerkung 5.4.4

Problem: Es ist nicht gesichert, dass die betrachtete Funktion die eben beschriebenen Bedingungen erfüllt. Da keine Startlösung bekannt ist, muss eine angenommen wer-

den. Diese kann außerhalb eines Bereiches $I =]a;b[$ liegen. Daher wird im nächsten Abschnitt eine Funktionsapproximation eingeführt, welche das Problem behebt.

Erweiterung auf mehrdimensionalen Input

Für den mehrdimensionalen Input erfolgen sämtliche Überlegungen analog, nur liegt die Betrachtungsweise eine Ableitungsstufe höher. Für die lokale Approximation der Log-Likelihood-Funktion wird ihre **Taylorreihenentwicklung zweiter Ordnung** verwendet. Für eine kurze Einführung siehe [TAY]. Die Annäherung an die Log-Likelihood-Funktion an der Stelle $\boldsymbol{\beta}$ erfolgt durch die Taylorreihenentwicklung an der Stelle $\boldsymbol{\beta}^{(t)}$ wie in Abbildung 5.4 zu sehen ist

$$L(\boldsymbol{\beta}) \approx L_{Taylor}(\boldsymbol{\beta}^{(t)}) = L(\boldsymbol{\beta}^{(t)}) + \left(\boldsymbol{u}^{(t)}\right)^{\mathsf{T}} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^{\mathsf{T}} \boldsymbol{H}^{(t)} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}). \tag{5.38}$$

Für diese Funktion wird nun die Ableitung gebildet, bei der die Nullstelle der entstehenden Gleichung

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx \boldsymbol{u}^{(t)} + \boldsymbol{H}^{(t)} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) = 0$$

zu ermitteln ist.

Durch Umstellung ergibt sich folgende Rekursionsvorschrift für die Berechnung des Maximums der taylorreihenapproximierten Maximum-Likelihood-Funktion, deren Wirkungsweise im obigen Abschnitt bereits für den eindimensionalen Fall bildlich erklärt wurde

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\boldsymbol{H}^{(t)}\right)^{-1} \boldsymbol{u}^{(t)}. \tag{5.39}$$

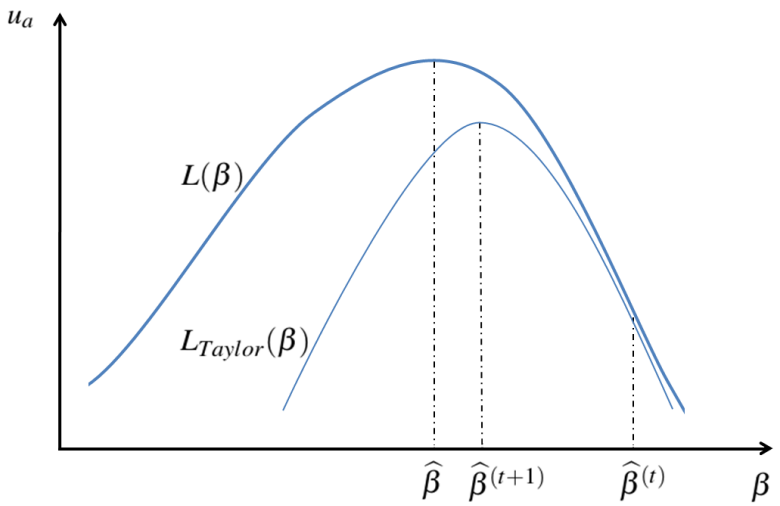


Abbildung 5.4: Darstellung des Ergebnisses der Anwendung des Newton-Raphson-Verfahrens auf die taylorreihenapproximierten Maximum-Likelihood-Funktion. $L(\beta)$ wird an der Stelle $\widehat{\beta}^{(t)}$ durch die Taylorreihe $L_{Taylor}(\beta^{(t)})$ approximiert. Mittels des Newton-Raphson-Verfahrens wird das Maximum der Taylorreihe an der Stelle $\widehat{\beta}^{(t+1)}$ bestimmt. Die Iteration wird mit der Stelle $\widehat{\beta}^{(t+1)}$ fortgesetzt.

Abbruchkriterien

Für Rekursionsvorschriften existieren oft vielerlei Kriterien, die zum Abbruch führen können. Hier erfolgt der Stop des Newton-Raphson-Verfahrens nach einer endlichen Anzahl von Iterationsschritten 6600 oder im Falle, dass die Änderung des Schätzparameters klein wird $\|\widehat{\boldsymbol{\beta}}^{(t+1)} - \widehat{\boldsymbol{\beta}}^{(t)}\| < \varepsilon$ mit $\varepsilon = 10^{-5}$.
Für beide Fälle kann das Abbruchkriterium zu einem Zeitpunkt erfüllt sein, an dem eine weit bessere Lösung unmittelbar bevorgestanden hätte. Ihre konkreten Realisierungen bilden einen Kompromiss aus Zeit und Güte. Vor allem die Differenz der $\boldsymbol{\beta}^{(t)}$ ist ausschlaggebend für die benötigte Zeit der Rekursion. Für $\varepsilon = 10^{-6}$ standen die nötigte Zeitdauer und die erreichte Näherung oft in keinerlei günstigen Relation zueinander. Häufig war die Güte der Näherung bereits vergleichbar mit der von $\varepsilon = 10^{-5}$.

Bemerkung 5.4.5
Es können nicht alle Wege berechnet werden. Zum Einen sind die dazu erforderlichen Datensätze nicht vollständig genug und zum Anderen gibt es auch Abbrüche der Rekursion. Oftmals ist dies eine Frage der Einstellung der Modellparameter. Werden Einstellungen aufgeweicht, bringt dies meist eine geringere Güte des Ergebnisses mit sich.

Ermittlung einer Näherungslösung für die Inverse der Hessematrix mittels Newton-konjugiertem Gradientenabstieg

Für die Berechnung der invertierten Hessematrix existieren einige Methoden, wie das Jacobi-, das Gauß-Seidel- oder das symmetrische Gauß-Seidel-Verfahren. Hier vorgestellt und angewendet werden soll aber das 1952 von Hestenes und Stiefel [HÜF06] vorgestellte Verfahren des Newton-konjugierten Gradientenabstiegs. Für die Lösung linearer Gleichungssysteme der Form $\boldsymbol{Ax} = \boldsymbol{b}$ liefert der unten aufgeführte Algorithmus das vermutlich effizienteste Iterationsverfahren. Der Abschnitt orientiert sich dabei am Buch [HAN09]. Es ist allerdings auch eines der wenigen effizienten Verfahren zur Lösung hochdimensionaler nichtlinearer Probleme ohne Nebenbedingungen, was 1964 von Fletcher und Reeves [FLE64] gezeigt wurde .

Vorbereitung

Zu Beginn muss die aus dem vorherigen Abschnitten gewonnene Gleichung (5.39)

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\boldsymbol{H}^{(t)}\right)^{-1} \boldsymbol{u}^{(t)}$$

in die Form eines linearen Gleichungssystems gebracht werden. Durch einfache Multiplikation mit der Hessematrix ergibt sich

$$\boldsymbol{H}^{(t)}\boldsymbol{\beta}^{(t+1)} = \boldsymbol{H}^{(t)}\boldsymbol{\beta}^{(t)} - \boldsymbol{u}^{(t)}. \tag{5.40}$$

Diese Umformung ist zulässig, da die Gleichung zur Bestimmung von $\boldsymbol{\beta}^{(t+1)}$ selbst aus einer Umstellung hervorging, welche die Existenz der inversen Hessematrix voraussetzte. Es erfolgt damit also eine Rückführung auf bereits bekannte und zulässige Werte. Nun liegt das gewünschte Gleichungssystem der Form $\boldsymbol{Ax} = \boldsymbol{b}$ vor. Dabei gilt

$$\boldsymbol{x} = \boldsymbol{\beta}^{(t+1)} \tag{5.41}$$

$$\boldsymbol{A} = \boldsymbol{H}^{(t)} \tag{5.42}$$

$$\boldsymbol{b} = \boldsymbol{H}^{(t)}\boldsymbol{\beta}^{(t)} - \boldsymbol{u}^{(t)} \tag{5.43}$$

$$\widehat{\boldsymbol{x}} = \boldsymbol{A}^{-1}\boldsymbol{b} = \widehat{\boldsymbol{\beta}}^{(t+1)}. \tag{5.44}$$

Zur besseren Lesbarkeit werden diese Termini im weiteren Verlauf des Abschnittes verwendet.

Die erforderliche Grundvoraussetzung zur Lösung dieses Systems sind die Eigenschaften der Symmetrie und der positiven Definitheit bei der Koeffizientenmatrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$. Da es sich bei dieser um die Hessematrix handelt, ist die Symmetriebedingung erfüllt. Ansonsten ist dies anhand von Formel (5.32) zu sehen. Die positive Definitheit der Hessematrix ist also nachzuweisen. Der Versuch eines Nachweises über den in der entsprechenden Fachliteratur verwendeten Weg führt zu einem Ringschluss. Für die Anwendung ist daher einfach vorauszusetzen, dass der Algorithmus dennoch funktioniert. Im Allgemeinen gelingt die Berechnung der Inversen. Die Theorie zeigt aber, dass dies nicht sein muss.

Zur Lösung des linearen Gleichungssystems wird die quadratische Form benötigt

$$\Phi(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Ax} - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{b}. \tag{5.45}$$

Die Betrachtung von $\boldsymbol{x}$ und seiner geschätzten Lösung $\widehat{\boldsymbol{x}}$ liefert

$$\begin{aligned}\Phi(\boldsymbol{x}) - \Phi(\widehat{\boldsymbol{x}}) &= \frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Ax} - \boldsymbol{x}^{\mathsf{T}} - \frac{1}{2}\widehat{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{A}\widehat{\boldsymbol{x}} + \widehat{\boldsymbol{x}}^{\mathsf{T}} \\ &= \frac{1}{2}(\boldsymbol{x}-\widehat{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{A}(\boldsymbol{x}-\widehat{\boldsymbol{x}}) + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{b} + \widehat{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{b} \\ &= \frac{1}{2}(\boldsymbol{x}-\widehat{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{A}(\boldsymbol{x}-\widehat{\boldsymbol{x}}) \geq 0.\end{aligned} \tag{5.46}$$

Mit der positiven Definitheit von $\boldsymbol{A}$ folgt die Nichtnegativität des letzten Terms. Er ist genau dann 0, wenn $\boldsymbol{x} = \widehat{\boldsymbol{x}}$ gilt. Damit entspricht das globale Minimum des Funktionals (5.45) der eindeutigen Lösung des linearen Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$.

Im Folgenden wird unten stehende Funktion benötigt:

Definition 5.4.4 (Energienorm, Innenprodukt)
Sei $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ symmetrisch und positiv definit, dann wird durch

$$\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}} \qquad \boldsymbol{x} \in \mathbb{R}^p \tag{5.47}$$

*eine Norm in $\mathbb{R}^p$ definiert, die sogenannte **Energienorm**. Zu dieser Energienorm sei das **Innenprodukt***

$$\langle \boldsymbol{x}, \boldsymbol{z}\rangle_{\boldsymbol{A}} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{z} \qquad \boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^p \tag{5.48}$$

definiert.

Diese Definitionen sind notwendig für die graphische Veranschaulichung des Verfahrens. Dazu sei die Abweichung (5.46) des Funktionals Φ von seinem Minimum erweitert

$$\Phi(\boldsymbol{x}) - \Phi(\widehat{\boldsymbol{x}}) = \frac{1}{2}(\boldsymbol{x}-\widehat{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{A}(\boldsymbol{x}-\widehat{\boldsymbol{x}}) = \frac{1}{2}\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|_{\boldsymbol{A}}^2. \tag{5.49}$$

Damit ist für den Abstand von $\boldsymbol{x}$ und $\widehat{\boldsymbol{x}}$ ein geeignetes Fehlermaß gewonnen. Geometrisch betrachtet bildet $\widehat{\boldsymbol{x}}$ dabei den Mittelpunkt einer Kugel bezüglich der Energienorm.

Ermittlung der Verfahrensvorschriften
Nachdem die einleitenden Betrachtungen und Definitionen geklärt sind, kann das Hauptproblem behandelt werden: Die Bestimmung der Iterationsvorschriften zur Approximation von $\widehat{\boldsymbol{x}}$.
Da bei der Iteration das globale Minimum einer quadratischen Form gefunden werden soll, führt die sukzessive Minimierung von Φ zu eben jenem Minimum.

Die Idee dazu ist folgende:
Sei $\boldsymbol{x}^{(k)}$ die aktuelle Iteration. Es wird eine neue „Suchrichtung“ $\boldsymbol{d}^{(k)} \neq 0$ bestimmt, um $\boldsymbol{x}^{(k+1)}$ mittels

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha\boldsymbol{d}^{(k)} \tag{5.50}$$

zu berechnen.

Eingesetzt in Φ ergibt sich ein Funktionalwert in Abhängigkeit von α

$$\Phi\left(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{d}^{(k)}\right) = \Phi\left(\boldsymbol{x}^{(k)}\right) + \alpha\boldsymbol{d}^{(k)}\boldsymbol{A}\boldsymbol{x}^{(k)} + \frac{1}{2}\alpha^2\boldsymbol{d}^{(k)^{\mathsf{T}}}\boldsymbol{A}\boldsymbol{d}^{(k)} - \alpha\boldsymbol{d}^{(k)^{\mathsf{T}}}\boldsymbol{b}. \tag{5.51}$$

Gegeben dem Fall, dass $\boldsymbol{x}^{(k)}$ und $\boldsymbol{d}^{(k)}$ bereits bekannt sind, erfolgt eine Minimierung des Funktionalwertes über die Anpassung von α, welche durch die Differentiation nach α

gewonnen wird. Es ergibt sich damit

$$\alpha_k = \frac{\boldsymbol{r}^{(k)\mathsf{T}}\boldsymbol{d}^{(k)}}{\boldsymbol{d}^{(k)\mathsf{T}}\boldsymbol{A}\boldsymbol{d}^{(k)}} \tag{5.52}$$

$$\text{mit } \boldsymbol{r}^{(k)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)} \text{ dem Residuum des Schrittes } k. \tag{5.53}$$

Aufgrund der positiven Definitheit von $\boldsymbol{A}$ ist der Nenner ungleich 0.

Noch nicht geklärt ist die Bestimmung der bis dato unbekannten Suchrichtung $\boldsymbol{d}^{(k)}$. Diese ergibt sich nach einfacher Rechnung aus der Darstellung (5.51) von $\Phi\left(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{d}^{(k)}\right)$

$$\begin{aligned}
\frac{\partial\Phi\left(\boldsymbol{x}^{(k)}\right)}{\partial\boldsymbol{d}^{(k)}} &= \text{grad}\Phi\left(\boldsymbol{x}^{(k)}\right)^{\mathsf{T}}\boldsymbol{d}^{(k)} \\
&= \lim_{\alpha\to 0}\frac{\Phi\left(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{d}^{(k)}\right) - \Phi\left(\boldsymbol{x}^{(k)}\right)}{\alpha} \\
&= \boldsymbol{d}^{(k)\mathsf{T}}\left(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b}\right).
\end{aligned}$$

Bei Betrachtung der Faktoren von $\boldsymbol{d}^{(k)}$ der ersten und der letzten Zeile folgt, dass

$$\text{grad}\Phi(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \tag{5.54}$$

der Gradient von Φ in $\boldsymbol{x}$ ist. Es ist zu sehen, dass das Residuum und der negative Gradient, welcher die Richtung des steilsten Abstiegs von Φ angibt, identisch sind. Das Residuum allein ist leider nicht immer die beste Entscheidung für die Suchrichtung.

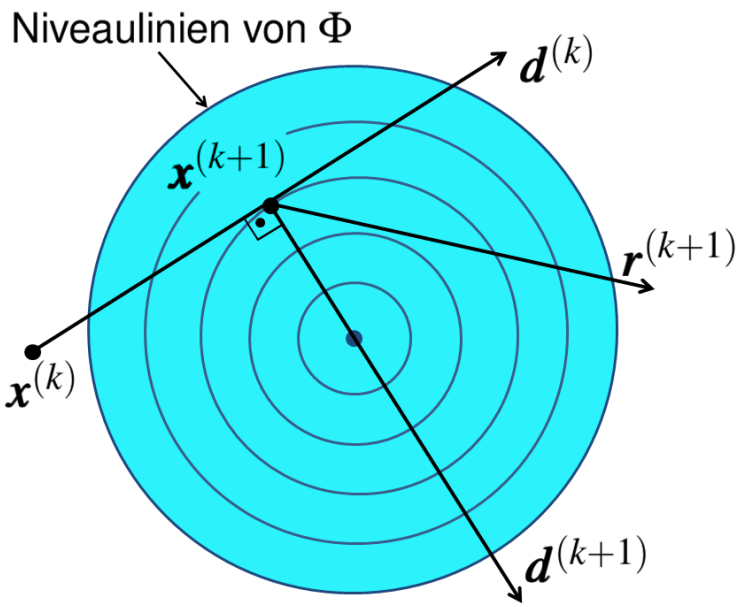


Abbildung 5.5: Skizze der $\|\cdot\|_{\boldsymbol{A}}$-Geometrie: Optimale Suchrichtung ist $\boldsymbol{d}^{(k+1)}$. Die Abbildung ist [HAN09] nachempfunden.

Eine Verdeutlichung der geometrischen Bedeutung der Aussagen (5.50), (5.52), (5.53) und (5.54) und der durch die Energienorm erzeugten Geometrie liefert Abbildung 5.5. Dabei spannen $\boldsymbol{d}^{(k)}$ und $\boldsymbol{r}^{(k+1)}$ eine Ebene auf. In der Abbildung sind die Höhenlinien von Φ in dieser Ebene zu sehen. Dabei sind die Niveauflächen von Φ gerade Kugeloberflächen deren Zentrum die Approximation $\widehat{\boldsymbol{x}}$ bildet. Die konzentrischen Kreise sind also die Niveaulinien der abgebildeten Ebene mit einem gemeinsamen Mittelpunkt. Dieser Mittelpunkt ist die Minimalstelle von Φ über dieser Ebene.
Nun existiert der Vektor $\boldsymbol{d}^{(k+1)}$ in dieser Ebene. Er steht senkrecht zu $\boldsymbol{d}^{(k)}$ und bietet sich als neue Suchrichtung an.
Es ist zu bemerken, dass eine Höhenlinie im Punkt $\boldsymbol{x}^{(k+1)}$ von der Gerade $\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}$, $\alpha \in \mathbb{R}$ berührt wird. Die Minimaleigenschaft des zugehöriges Parameters α_k führt zu diesem Ergebnis. Wie in der Abbildung zu sehen ist, führt die richtige Wahl der nachfolgenden Schrittweite α_{k+1} aufgrund der Gleichung (5.53) beim nächsten Schritt automatisch in den Mittelpunkt des Kreises.

Aus diesen Überlegungen folgt die neue Suchrichtung

$$\boldsymbol{d}^{(k+1)} = \boldsymbol{r}^{(k+1)} + \beta_k \boldsymbol{d}^{(k)} \quad \text{mit} \quad \langle \boldsymbol{d}^{(k+1)}, \boldsymbol{d}^{(k)} \rangle_{\boldsymbol{A}} = 0. \tag{5.55}$$

Damit berechnet sich die neue Schrittweite β_k über

$$\beta_k = -\frac{\boldsymbol{r}^{(k+1)^{\mathsf{T}}} \boldsymbol{A} \boldsymbol{d}^{(k)}}{\boldsymbol{d}^{(k)^{\mathsf{T}}} \boldsymbol{A} \boldsymbol{d}^{(k)}}. \tag{5.56}$$

Eine Wohldefiniertheit für die Gleichungen (5.53) und (5.56) liegt nur vor, wenn $\boldsymbol{d}^{(k)} \neq \boldsymbol{0}$ gilt. Dieser Fall kann laut (5.55) allerdings nur im Falle der linearen Abhängigkeit von $\boldsymbol{r}^{(k)}$ und $\boldsymbol{d}^{(k-1)}$ eintreten.
$\boldsymbol{d}^{(k-1)}$ liegt jedoch tangential zu der Niveaufläche von Φ. Der Gradient $\boldsymbol{r}^{(k)}$ liegt bezüglich des euklidischen Innenprodukts also orthogonal zu $\boldsymbol{d}^{(k-1)}$. Die lineare Abhängigkeit gilt somit nur für $\boldsymbol{r}^{(k)} = \boldsymbol{0}$, was mit dem Fall gleichzusetzen ist, dass die Lösung mit $\boldsymbol{x}^{(k)} = \widehat{\boldsymbol{x}}$ gefunden wurde.
Für $\boldsymbol{x}^{(k)} \neq \widehat{\boldsymbol{x}}$ liegt damit ein durch die Anweisungen (5.50) bis (5.56) erzeugter wohldefinierter Algorithmus vor.

Die spezielle Orthogonalitätsbedingung der Suchrichtungen aus (5.55),
$\langle \boldsymbol{d}^{(k+1)}, \boldsymbol{d}^{(k)} \rangle_{\boldsymbol{A}} = 0$, wird **zueinander A-konjugiert** genannt. Sie ist die namensgebende Bedingung für das **Verfahren der Newton-konjugierten Gradienten**.

Ein entscheidender Punkt für die Güte dieses Verfahrens liefert eine Optimalitätseigenschaft (siehe nachfolgend Satz 5.4.1)

Lemma 5.4.1
Gegeben sei der beliebige Startvektor $\boldsymbol{x}^{(0)}$ *und* $\boldsymbol{d}^{(0)} = \boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$. *Sei* $\boldsymbol{x}^{(k)} \neq \widehat{\boldsymbol{x}}$ *für* $k = 0, ..., m$, *dann gilt*

$$
\begin{aligned}
(a) \qquad & \boldsymbol{r}^{(m)^T}\boldsymbol{d}^{(j)} = 0 \quad \forall\, 0 \leq j < m, \\
(b) \qquad & \boldsymbol{r}^{(m)^T}\boldsymbol{r}^{(j)} = 0 \quad \forall\, 0 \leq j < m, \\
(c) \qquad & \langle \boldsymbol{d}^{(m)}, \boldsymbol{d}^{(j)} \rangle_{\boldsymbol{A}} = 0 \quad \forall\, 0 \leq j < m.
\end{aligned}
$$

Der Beweis erfolgt über vollständige Induktion und wird an dieser Stelle nicht ausgeführt. Es sei aber auf [HAN09] Kapitel 9 Seite 90 verwiesen.

Eine wichtige Erkenntnis des Beweises ist die Gleichung

$$\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k \boldsymbol{A}\boldsymbol{d}^{(k)}. \tag{5.57}$$

Teil (c) des Lemmas lässt schließen, dass für alle Suchrichtungen gilt, dass sie paarweise $\boldsymbol{A}$-konjugiert sind. Nach (b) sind die Residuen allesamt linear unabhängig (da alle Orthogonalsysteme linear unabhängig sind). Nach maximal n Iterationsschritten gilt $\boldsymbol{r}^{(n)} = 0$ und somit $\boldsymbol{x}^{(n)} = \widehat{\boldsymbol{x}}$.

Korollar 5.4.1
Für $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ *symmetrisch und positiv definit findet das Newton-konjugierte Gradientenverfahren nach höchsten* n *Schritten die exakte Lösung* $\boldsymbol{x}^{(n)} = \widehat{\boldsymbol{x}}$.

Leider existiert eine nicht überraschende Diskrepanz bezüglich der Anwendbarkeit in Theorie und Praxis, die auf der Rundung numerischer Werte basiert. Dabei gehen bei zunehmender Iterationsdauer oft die Orthogonalitätseigenschaften aus Lemma 5.4.1 verlustig. Folglich verliert Korollar 5.4.1 seine Praxisrelevanz.

Die nächste Optimalitätseigenschaft ist allerdings von wesentlicher größerer Bedeutung für das Newton-konjugierte Gradientenverfahren als iteratives Verfahren.

Definition 5.4.5
Es sei $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ *und* $\boldsymbol{z} \in \mathbb{R}^p$. *Dann heißt der Teilvektorraum*

$$\mathrm{K}_k(\boldsymbol{A}, \boldsymbol{z}) = span\{\boldsymbol{z}, \boldsymbol{A}\boldsymbol{z}, ..., \boldsymbol{A}^{k-1}\boldsymbol{z}\} \tag{5.58}$$

Krylow-Raum *der Dimension* k *von* $\boldsymbol{A}$ *bezüglich* $\boldsymbol{z}$.

Satz 5.4.1
Es sei $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ symmetrisch und positiv definit. $\boldsymbol{x}^{(k)} \neq \widehat{\boldsymbol{x}}$ sei die k-te Iterierte des Newton-konjugierten Gradientenabstiegverfahrens. Weiterhin sei $\boldsymbol{d}^{(0)} = \boldsymbol{r}^{(0)}$. Dann gilt

$$\boldsymbol{x}^{(k)} \in \boldsymbol{x}^{(0)} + \mathrm{K}_k\left(\boldsymbol{A}, \boldsymbol{r}^{(0)}\right). \tag{5.59}$$

Dabei ist $\boldsymbol{x}^{(k)}$ die eindeutige Minimalstelle der Zielfunktion Φ in diesem affinen Raum.

Wie beim Beweis des Lemmas 5.4.1 wird der Nachweis nicht wiedergegeben. Er ist allerdings bei (Referenz NCGV 2009 ab 85 und 64) in Kapitel 9 auf Seite 92 nachzulesen.

Aufgrund der Bedeutung des Satzes werden dennoch die wichtigsten Ergebnisse zusammengefasst:

Mittels wiederholter vollständiger Induktion werden folgende Zwischenschritte gezeigt

$$\boldsymbol{d}^{(j)} \in \mathrm{span}\{\boldsymbol{r}^{(0)},\ ..,\ \boldsymbol{r}^{(j)}\} \qquad j = 0,\ ...,\ k-1.$$

Damit folgt

$$\mathrm{span}\{\boldsymbol{d}^{(0)},\ ...,\ \boldsymbol{d}^{(k-1)}\} = \mathrm{span}\{\boldsymbol{r}^{(0)},\ ...,\ \boldsymbol{r}^{(k-1)}\}.$$

Eine Folge aus (5.50) liefert

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(0)} + \sum_{j=0}^{k-1} \alpha_j \boldsymbol{d}^{(j)} \in \boldsymbol{x}^{(0)} + \mathrm{span}\{\boldsymbol{r}^{(0)},\ ...,\ \boldsymbol{r}^{(k-1)}\}.$$

Zu zeigen ist also die Zugehörigkeit der $\boldsymbol{r}^{(j)}$ zum aufspannenden Teilraum. Induktiv wird gezeigt, dass

$$\boldsymbol{r}^{(j)} \in \mathrm{span}\{\boldsymbol{r}^{(0)},\ ...,\ \boldsymbol{A}^j\boldsymbol{r}^{(0)}\} \qquad j = 0,\ ...,\ k-1$$

und somit

$$\mathrm{span}\{\boldsymbol{r}^{(0)},...,\boldsymbol{r}^{(k-1)}\} \subset \mathrm{span}\{\boldsymbol{r}^{(0)},\boldsymbol{A}\boldsymbol{r}^{(0)},...,\boldsymbol{A}^{k-1}\boldsymbol{r}^{(0)}\}.$$

Als Folge dieser Zwischenschritte ist abzuleiten, dass

$$\mathrm{span}\{\boldsymbol{d}^{(0)},\ ...,\ \boldsymbol{d}^{(k-1)}\} = \mathrm{span}\{\boldsymbol{r}^{(0)},\ ...,\ \boldsymbol{r}^{(k-1)}\} = K_k\left(\boldsymbol{A}, \boldsymbol{r}^{(0)}\right).$$

Mittels Korollar 5.4.1 und dem Zwischenergebnis lässt sich folgende Darstellung der Differenz zwischen einem beliebigen Element $\boldsymbol{x} \in \boldsymbol{x}^{(0)} + K_k\left(\boldsymbol{A}, \boldsymbol{r}^{(0)}\right)$ und dem approxi-

mierten Minimum herstellen

$$\begin{aligned}\widehat{\boldsymbol{x}}-\boldsymbol{x} &= \widehat{\boldsymbol{x}}-\boldsymbol{x}^{(k)}+\boldsymbol{x}^{(k)}-\boldsymbol{x}\\ &= \widehat{\boldsymbol{x}}-\boldsymbol{x}^{(k)}+\sum_{j=0}^{k-1}\delta_j\boldsymbol{d}^{(j)}\\ &= \sum_{j=0}^{k-1}\delta_j\boldsymbol{d}^{(j)}+\sum_{j=k}^{m-1}\alpha_j\boldsymbol{d}^{(j)} \text{ für gewisse } \delta_j\in\mathbb{R}.\end{aligned}$$

Durch die $\boldsymbol{A}$-konjugiertheit der Suchrichtungen nach Lemma 5.4.1(c) lässt sich der Satz des Pythagoras anwenden

$$\begin{aligned}\Phi(\boldsymbol{x})-\Phi(\widehat{\boldsymbol{x}}) &= \frac{1}{2}\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|\\ &= \frac{1}{2}\|\boldsymbol{x}^{(k)}-\widehat{\boldsymbol{x}}\|_{\boldsymbol{A}}^2+\frac{1}{2}\|\sum_{j=0}^{k-1}\delta_j\boldsymbol{d}^{(j)}\|_{\boldsymbol{A}}^2\\ &= \Phi(\boldsymbol{x}^{(k)})-\Phi(\widehat{\boldsymbol{x}})+\frac{1}{2}\|\sum_{j=0}^{k-1}\delta_j\boldsymbol{d}^{(j)}\|_{\boldsymbol{A}}^2.\end{aligned}$$

Es folgt, dass $\Phi(\boldsymbol{x})\geq\Phi(\boldsymbol{x}^{(k)})$ mit Gleichheit, genau dann, wenn $\boldsymbol{x}=\boldsymbol{x}^{(k)}$.

Ende des Nachweises.

In der Literatur wird stets darauf hingewiesen, dass anstelle der Gleichungen (5.53) und (5.56) die nachfolgend angegebenen numerisch stabileren Gleichungen (5.60) und (5.61) verwendet werden sollten.
Wichtig dafür ist eine aus Lemma 5.4.1(a) und (5.55) gewonnene Tatsache

$$\boldsymbol{r}^{(k)^{\mathsf{T}}}\boldsymbol{d}^{(k)}=\boldsymbol{r}^{(k)^{\mathsf{T}}}\boldsymbol{r}^{(k)}+\beta_{k-1}\boldsymbol{r}^{(k)^{\mathsf{T}}}\boldsymbol{d}^{(k-1)}=\boldsymbol{r}^{(k)^{\mathsf{T}}}\boldsymbol{r}^{(k)}.$$

Eingesetzt in (5.53) gilt damit

$$\alpha_k=\frac{\|\boldsymbol{r}^{(k)}\|_2^2}{\boldsymbol{d}^{(k)^{\mathsf{T}}}\boldsymbol{A}\boldsymbol{d}^{(k)}}. \tag{5.60}$$

Weiterhin folgt aufgrund von Lemma 5.4.1(b), (5.57) und (5.60)

$$\begin{aligned}\boldsymbol{r}^{(k+1)^{\mathsf{T}}}\boldsymbol{A}\boldsymbol{d}^{(k)} &= \frac{1}{\alpha_k}\left(\boldsymbol{r}^{(k+1)^{\mathsf{T}}}\boldsymbol{r}^{(k)}-\boldsymbol{r}^{(k+1)^{\mathsf{T}}}\boldsymbol{r}^{(k+1)}\right)\\ &= -\frac{1}{\alpha_k}\|\boldsymbol{r}^{(k+1)}\|_2^2\\ &= -\frac{\|\boldsymbol{r}^{(k+1)}\|_2^2}{\|\boldsymbol{r}^{(k)}\|_2^2}\boldsymbol{d}^{(k)^{\mathsf{T}}}\boldsymbol{A}\boldsymbol{d}^{(k)}.\end{aligned}$$

Damit gilt

$$\beta_k = \frac{\|\boldsymbol{r}^{(k+1)}\|_2^2}{\|\boldsymbol{r}^{(k)}\|_2^2}. \tag{5.61}$$

Der Algorithmus des Newton-konjugierten Gradientenverfahrens
Der vervollständigte Algorithmus kann nun angegeben werden.

Algorithm 1 Newton-konjugiertes Gradientenverfahren

Initialisierung:
$\boldsymbol{A} \in \mathbb{R}^{p\times p}$ sei symmetrisch und positiv definit
Wähle beliebiges $\boldsymbol{x}^{(0)} \in \mathbb{R}^p$
$\boldsymbol{r}^0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$
$\boldsymbol{d}^{(0)} = \boldsymbol{r}^{(0)}$
$k = 0$

while $\|\boldsymbol{r}^{(k)}\| > \min\left(0.5, \left(\mathbf{1}^\mathsf{T}\boldsymbol{b}\right)^2\right) \cdot \mathbf{1}^\mathsf{T}\boldsymbol{b}$ **do**
$\alpha_k = \dfrac{\|\boldsymbol{r}^{(k)}\|_2^2}{\boldsymbol{d}^{(k)^\mathsf{T}}\boldsymbol{A}\boldsymbol{d}^{(k)}}$
$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}$
$\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k \boldsymbol{A}\boldsymbol{d}^{(k)}$
$\beta_k = \dfrac{\|\boldsymbol{r}^{(k+1)}\|_2^2}{\|\boldsymbol{r}^{(k)}\|_2^2}$
$\boldsymbol{d}^{(k+1)} = \boldsymbol{r}^{(k+1)} + \beta_k \boldsymbol{d}^{(k)}$
$k = k + 1$
end while

return $\boldsymbol{x}^{(k)}$ {$\boldsymbol{x}^{(k)}$ ist die Approximation von $\boldsymbol{A}^{-1}\boldsymbol{b}$, $\boldsymbol{r}^{(k)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)}$ das zugehörige Residuum}

Bemerkung 5.4.6
Der Abbruchwert entstammt einem Programm, welches eine Variante des Newton-konjugierten Gradientenverfahrens als Funktion enthält und hat sich als praktikabel erwiesen [GAB].

Bemerkung 5.4.7
Es existieren verschiedene Varianten des Algorithmus, die sich hauptsächlich in der Berechnung des β_k unterscheiden. So zum Beispiel die Fletcher-Reeves- oder die Polak-Ribire-Variante, siehe [HÜF06].

Aufwand des Newton-konjugierten Gradientenverfahrens
Pro Iterationsschritt sind folgende Terme zu beachten:

1. Die Berechnung der Innenprodukte benötigt jeweils einen Aufwand von $O(p)$.

2. Es wird eine Matrix-Vektor-Multiplikation ($\boldsymbol{A}\boldsymbol{d}^{(k)}$) benötigt. Im vorliegenden Fall besitzt $\boldsymbol{A}$ wesentlich mehr als n von 0 verschiedene Einträge.

Damit liegt der Aufwand laut [HAN09] in etwa bei dem des Gesamt- und Einzelschrittverfahrens mit $O(p^2)$.
Liegt die Startlösung $\boldsymbol{x}^0$ in der Nähe von der wahren Lösung $\boldsymbol{x}$ so werden für vollbesetzte Systeme meist weniger als $2p$ Schritte benötigt.

Bemerkung 5.4.8
Im Zusammenhang mit dem Aufwand sei der Begriff der Präkonditionierung kurz angesprochen. Der Iterationsfehler des Newton-konjugierten Gradientenverfahrens kann mit der oberen Schranke $O(\left(1-2\text{cond}_2^{-\frac{1}{2}(\boldsymbol{A})}\right)^k)$ abgeschätzt werden. Damit folgt, dass eine bessere Konditionierung von $\boldsymbol{A}$ in der Regel eine schnellere Konvergenz des Verfahrens bedeutet. Dies wird durch die Erstellung einer **Präkonditionierungsmatrix $\boldsymbol{M}$** erreicht. Diese ist mit dem Gleichungssystem zu multiplizieren $(\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{M}^{-1}\boldsymbol{b})$, welches anschließend in der leicht abgewandelter Art und Weise gelöst wird. Die Implementierung übersteigt den Rahmen der Arbeit.

5.4.7 Erweiterung zur allgemeinen Schätzung

Das vorgestellte Verfahren der logistischen Regression kann den bereits in Abschnitt 5.2 auftretenden Nachteil der Datenarmut besitzen. Da für einen Weg alle Routen mehrerer Jahrgänge einbezogen werden, tritt dieser Fall lediglich bei Wegen mit sehr wenigen Routen oder bei der Verwendung weniger Jahrgänge auf. Das einzige Steuerungselement ist die Anzahl der verarbeiteten Jahrgänge, wobei das bereits erörterte Problem der Aktualität der Daten auftreten kann. Mit der Erstellung eines einzelnen $\boldsymbol{\beta}$ für alle Wege können beide Nachteile behoben werden.
Anstatt also die Informationen der Beobachtungen für einen Weg aus verschiedenen Jahren zu beziehen, wird ein einziges Jahr verwendet. Die Informationen werden für alle Routen aller Wege erstellt. Die Formeln sind $1:1$ zu übernehmen. Nur die Trennung in separate Wege bei der Informationsgewinnung entfällt. Es entsteht eine einzige große Matrix $\boldsymbol{X}$ (und $\boldsymbol{y}$, $\boldsymbol{n}$, etc.). Das Ergebnis bildet ein allgemeines $\boldsymbol{\beta}_{\text{ges}}$, das für alle Routen anwendbar ist. Es kann auch als Startwert für die Berechnungen der wegspezifischen $\boldsymbol{\beta}$ dienen.
Weitere Vorteile dieses Verfahrens ist die erhöhte Stabilität gegen Ausreißer, die Vermeidung der Überanpassung und die erhöhte Geschwindigkeit, da die Daten aus lediglich einem Jahr zu bearbeiten sind. Nachteilig allein ist die zu erwartende erhöhte

Abweichung im Vergleich zu den spezifizierten $\boldsymbol{\beta}$.

5.4.8 Vor- und Nachteile

Wie eingangs des Abschnittes bereits gesagt, liegt der Hauptvorteil dieses Verfahrens in der Einbeziehung mehrerer Eigenschaften einer Route, sodass ein kausaler Zusammenhang zwischen Einflussfaktoren und der Auswahl einer Route erstellt werden kann. Bisher gab es lediglich die Möglichkeit, die Anzahl der Personen beziehungsweise ihre relative Häufigkeiten zu verwenden. Mit diesem Modell wird weiteren wichtigen Eigenschaftsparametern die Möglichkeit geboten, in die Bewertung eingehen zu können. Welche Parameter wichtig sind und welche nicht, kann durch eine anschließende Analyse der $\boldsymbol{\beta}$-Werte bestimmt werden. Das Modell ist entsprechend erweiterbar.
Weiterhin kann das Verfahren zu einer allgemeinen Schätzung erweitert werden. Siehe Abschnitt 5.4.7. Die darin bereits aufgeführten vermeidbaren Vor- und Nachteile werden aufgrund der Nähe des Abschnittes nicht erneut aufgelistet.

Wegen der Komplexität des Modells sind die $\boldsymbol{\beta}$ nur bedingt für eine direkte Interpretation der Bedeutung der Einflussfaktoren geeignet. Wie oben angesprochen sind weitere Analysemodelle von Nöten, um die Relevanz der einzelnen Inputparameter zu bestimmen. Aussagen über direkte Korrelation und Bedeutung der Größe der Werte lassen sich nur schlecht treffen.
Ein Nachteil dieses Verfahrens ist seine Geschwindigkeit die aufgrund der zwei geschachtelten Iterationen geringer ausfällt als bei den vorherigen Modellen. Für diese Schachtelung existiert eine Implementation nach [GAB], sodass es bei dessen Anwendung in der Praxis eines vergleichsweise ebenso geringen Zeitaufwandes bedarf wie bei Modell 3. Bei entsprechender Parameterwahl ist der Zeitbedarf von Modell 3 sogar größer.
Der Hauptnachteil des Modells liegt in den Anwendungsvoraussetzungen, denn es ist nicht auf multiplen Output ausgelegt. Im vorliegenden Fall ist das System eher dafür da, um für *eine einzige Route* zu bestimmen, wie viele Leute dieses Route wählen beziehungsweise nicht wählen. Nach Rundung des Ergebnisses führt dies im Normalfall zu einem binären Output. Der nichtbinäre Output wird jedoch mit den nichtbinären Outputs der restlichen Routen des Weges kombiniert, wie in Abschnitt 5.4.5 beschrieben. Die nachfolgend aufgelisteten Nachteile sind die Folgen der Wahl dieses nicht geeigneten Systems.
Nachteil 3: Das Modell ist für binäre Entscheidungen ausgelegt. Die Outputs für die verschiedenen Routen eines zu schätzenden Beobachtungszeitpunktes t summieren sich nicht zu eins. Sie stehen also nicht in direkter Konkurrenz zueinander. Eine Route gewinnt also nicht an Bedeutung beim Ausfall oder Bedeutungsverlust einer anderen. Zumindest nicht in den Werten des Outputs. Die in Abschnitt 5.4.5 eingeführte Normierung der Werte bildet diese Bedeutungsverteilung weder im praktischen, noch im theoretischen Sinn adäquat ab.

Nachteil 4: Die nötige Unabhängigkeit der Daten ist nicht gegeben. Diese ist für einen bestimmten Beobachtungszeitpunkt nicht vorhanden, da die pure Existenz weiterer Routen den normierten Output beeinflusst. Dies gestaltet sich folgendermaßen: Eine Route a besitzt den Outputwert 0.3. Es existiert eine weitere Route b mit Outputwert 0.2. Nach Normierung würden 60% der Passagiere auf a und 40% auf b verteilt werden. Bei Existenz einer weiteren Route mit Outputwert 0.2 würden nun nach Normierung weniger Passagiere auf die Routen a und b verteilt werden. Somit hat bereits die Existenz einer weiteren Route die Verteilung beeinflusst. Dies gilt nur, wenn ein Passagier eine Route geflogen ist.
Nachteil 5: Wie in Formel (5.25) zu sehen und bei den vorhergehenden Nachteilen angesprochen ist, gibt es kein Zusammenspiel mehrerer Routen bei der Ermittlung der Auswahlwahrscheinlichkeit.

Im Gegensatz zu den Modellen 1, 2 und 3 sind diese Probleme von schwerwiegender Natur, da die Voraussetzungen und das Anwendungsgebiet eindeutig nicht für das vorliegende Problem geeignet sind und die relativen Häufigkeiten nicht direkt aus dem Modell ableitbar sind. Bei den vorherigen Modellen war dies nicht der Fall. Es stand lediglich die Frage im Raum, ob das Problem von linearer Natur ist oder nicht.
Modell 4 wird deswegen aufgeführt, weil es einerseits die Grundlage für Modell 5 bietet und andererseits implementiert wurde, um über einen Vergleichswert zu verfügen. Denn wie bereits in Kapitel 4 erwähnt, steigt die Instabilität eines Modells mit seiner Komplexität.
Die meisten dieser Probleme werden durch die Anwendung des Modelles 5, des **Conditional Logit**, behoben. Dabei handelt es sich um eine der vielen Erweiterungen der logistischen Regression, die sich auf multiple Outputs konzentriert. Es fließen mehrere Routen in die Berechnung ein und der Output wird gleichzeitig für verschiedene Routen ermittelt. Die Modellvoraussetzungen sind optimal für das Problem. Lediglich die Geschwindigkeit wird zugunsten der Komplexität weiter abnehmen und das Problem der Unabhängigkeit der Daten wird durch Verschiebung in die Berechnung nicht vollständig aufgehobenen, aber abgeschwächt sein.

5.5 Modell 5: Conditional Logit

Wie Ausgangs des Abschnittes 5.4.8 erklärt, wird in diesem Abschnitt eine Erweiterung der logistischen Regression behandelt: Der **Conditional Logit**. Es ist nun möglich, mehrere Inputparameter von mehreren Alternativen gleichzeitig in einer Formel zu verarbeiten. Die errechneten Wahrscheinlichkeiten summieren sich nun für die Routen eines Weges zu eins, sodass in diesem Fall von einer Auswahlwahrscheinlichkeit zu sprechen ist, welche sich an den Alternativrouten orientiert.
Sämtliche theoretische und praktische Hintergründe des Modells, des Verfahrens und der Schätzung von $\boldsymbol{\beta}$ wurden bereits im vorherigen Abschnitt behandelt. Da sich nichts am bereits beschriebenen Vorgehen ändert, sind in diesem Abschnitt lediglich die ge-

änderten Modellvoraussetzungen und Formeln anzuführen.
Herleitung und Argumentation sind an [AGR02] angelehnt.

Durch die Einbeziehung mehrerer Alternativen ändert sich die Interpretation einer Beobachtung. Ihre Elemente x und y besitzen nun zwei Indize i und j, wobei i für den Monat der Beobachtung steht und j für die Route eines bestimmten Weges.

Bemerkung 5.5.1
$\boldsymbol{x}$ (noch ohne Index) beinhaltet wieder alle in Abschnitt 3.2.2 (ohne x_{a1}, x_{a2} und x_{a3}) aufgeführten Werte einer Route als Inputdaten. Auch y besteht wieder aus der Anzahl der Passagiere einer Route.

5.5.1 Variablen

Der Conditional Logit kann wie bei der logistischen Regression für jeden Weg einzeln angewandt werden. Aus Gründen der Lesbarkeit wird daher auf einen Wegindex verzichtet. Sämtliche in Tabelle 5.5 eingeführten Variablen gelten für genau einen Weg.

Variable	Beschreibung
$\boldsymbol{\beta}$	$p \times 1$ - Vektor der zu schätzenden Koeffizienten
$\boldsymbol{\beta}^{(t)}$	$p \times 1$ - Vektor der zu schätzenden Koeffizienten zum Berechnungszeitpunkt t
p	Anzahl Inputparameter $=$ Anzahl Koeffizienten von $\boldsymbol{\beta}$
$N, (i = 1, ..., N)$	Anzahl der Beobachtungsmonate
$J_i, (j = 1, ..., J_i)$	Anzahl der Routen im Monat i
$\mathbb{X}$	Gesamten Inputs aller Monate
$\boldsymbol{X}_i$	$N \times p$ - Inputmatrix des Monates i
$\boldsymbol{x}_{ij}$	$1 \times p$ - Vektor, Zeile der Inputmatrix des Monates $i =$ Inputparameter einer Route j nach Abschnitt 3.2.2 (ohne x_{a1}, x_{a2} und x_{a3})
$x_{ija}, (a = 1, ..., p)$	Inputparameter a der Inputzeile x_{ij} (also der Route j des Monates i) der Inputmatrix $\boldsymbol{x}_i$
R_i	Gesamtzahl der Passagiere des Monates i
$\boldsymbol{y}_i$	$1 \times J_i$ - Passagiervektor des Monates i
y_{ij}	Passagiere der Route j im Monat i
$\pi_j(\boldsymbol{X}_i)$	Auswahlwahrscheinlichkeit einer Route j im Monat i
$\boldsymbol{u} = (u_a^{(t)}, (a = 1, ..., p))$	$p \times 1$ - Gradient der ersten Ableitung der logarithmischen Zielfunktion zum Berechnungszeitpunkt t
$\boldsymbol{H}^{(t)} = (h_{ab}^{(t)}, (a, b = 1, ..., p))$	$p \times p$ - Hessematrix der zweiten Ableitung der logarithmischen Zielfunktion zum Berechnungszeitpunkt t

Tabelle 5.5: Variablen dieses Abschnittes

5.5.2 Das Modell

Einführung

Der Conditional Logit zählt wie die logistische Regression zu den Regressionsmodellen. Darüber hinaus ist er der Unterklasse der **Discrete Choice Modelle** zuzuordnen. Eine Einführung dazu ist in [DAG00] gegeben. Diese entscheidungs- beziehungsweise auswahlbasierten Modelle dienen der Analyse von ökonomischen Präferenzen. Sie beschreiben, erklären und sagen die Entscheidungen zwischen zwei oder mehr Alternativen vorher, die von einem einzelnen Subjekt getroffen werden können. Die Auswahl zwischen den verschiedenen Routen eines Weges bilden diese Alternativen und werden durch eine Anzahl von Attributen beschrieben, welche das Subjekt zu bewerten hat und an Hand derer es sich entscheidet. Diese Attribute entsprechen den bereits bekannten Inputparametern einer Route. Hauptanwendungsgebiete sind Transport, Gesundheitsökonomik, Marketing und Umweltökonomik.

Zwei ökonomische Theorien liefern die Richtlinien für die Theorie der Discrete Choice Modelle [LAN08].
Zum Einen die Konsumtheorie des amerikanischen Ökonoms Kelvin Lancaster. Sie besagt, dass „Menschen ihren Nutzen nicht aus dem Konsum von abstrakten Gütern ziehen, sondern aus den Eigenschaften dieser Güter“ [LAN71].
Zum Anderen liefert der in Abschnitt 5.4.1 bereits erwähnte Daniel McFadden einen theoretischen Hintergrund mit seiner Random-Utility-Theorie in der „Menschen immer das Gut mit den ihnen zugänglichen Gütern auswählen, dem sie den größten Nutzen zuschreiben (Nutzenmaximierung)“ [LOU00]. Dieser Nutzen soll laut der Random-Utility-Theorie aus zwei Komponenten bestehen: Eine von den Eigenschaften abhängige, systematische Komponente und einer zufälligen Komponente. Letztere ist in der Mathematik der Fehler des weißen Rauschens. Mittels der Analyse der Auswahlen kann die systematische Komponente bestimmt werden, was der Anwendung des Conditional Logit auf die Routeneigenschaften entspricht. Je nach Annahme über die Verteilung der Fehlerterme werden verschiedene statistische Modelle benutzt.
Die typischen Modelle sind der **Probit**, der **Multinomial Logit**, dessen Erweiterung, der **Conditional Logit** und wiederum dessen Unterarten, der **Nested Logit** und der **Mixed Logit**.

Der Multinomial Logit

Es sei zuerst der Multinomial Logit betrachtet, welcher das Bindeglied zwischen der logistischen Regression und dem Conditional Logit herstellt. Dazu sei die Random-Utility-Theorie von McFadden realisiert, die von einer Nutzenfunktion mit einer systematischen und einer zufälligen Komponente spricht [PRI]. Es sei darauf hingewiesen, dass eine abweichende Indizierung zu den in Abschnitt 5.4.6 vorgestellten Variablen auftritt.

Definition 5.5.1 (Nutzenfunktion)
Sei Y_i die diskrete Wahl des i-ten Individuums für die j-te Alternative aus einer Menge von J_i Alternativen und sei U_{ij} der Wert dieser Entscheidung für Individuum i. Weiterhin sei η_{ij} eine systematischen Komponente und ε_{ij}, sodass U_{ij} eine unabhängige Zufallsvariable ist für die gilt

$$U_{ij} = \eta_{ij} + \varepsilon_{ij}. \tag{5.62}$$

*Dann heißt U_{ij} **Nutzenfunktion** eines Discrete Coice Modells.*

Wenn angenommen wird, dass menschliche Individuen bestrebt sind, den größtmöglichen Nutzen aus ihren Entscheidungen zu ziehen, dann wird ein Individuum i Alternative j wählen, wenn $U_{ij} \geq U_{i1}, ..., U_{iJ_i}$. Die entsprechende Wahrscheinlichkeit für diese Wahl kann angegeben werden mit

$$\pi_{ij} = P(Y_i = j) = P\left(\max(U_{i1}, ..., U_{iJ_i}) = U_{ij}\right) \tag{5.63}$$

Für das nachfolgende Lemma sei folgende Verteilung eingeführt.

Definition 5.5.2 (Gumbelverteilung) *Sei X eine stetige Zufallsvariable. Sei $\lambda > 0$ ein Skalierungsparameter und $\kappa \in \mathbb{R}$ ein Lageparameter. Dann ist die **Gumbelverteilung** (auch Typ-1-Extremwertverteilung beziehungsweise Extremal-1-Verteilung genannt) definiert durch die Dichtefunktion*

$$f(x) = \lambda e^{-\lambda(x-\kappa)} e^{-e^{-\lambda(x-\kappa)}}$$

und die Verteilungsfunktion

$$F(x) = e^{-e^{-\lambda(x-\kappa)}}.$$

Nach [MAI90] steht der Lageparameter κ für den Wert der höchsten Dichte (Modalwert). Die Inverse der Standardabweichung σ ist proportional zum Skalierungsparameter λ

$$\kappa = x_{\text{modal}}, \quad \lambda = \frac{\pi}{\sqrt{6}\sigma}.$$

Die Werte der Standard-Gumbelverteilung lauten $\lambda = 1$ und $\kappa = 0$.

Es gilt folgendes Lemma:

Lemma 5.5.1
Es gelte die Aufteilung der Nutzenfunktion von McFadden $U_j = \eta_j + \varepsilon_j$. Gegeben sei die zufällige Komponente η, der gumbelverteilte Fehlerterm ε mit der Dichte

$$f(\varepsilon) = e^{-(\varepsilon+\alpha)-e^{-(\varepsilon+\alpha)}} \tag{5.64}$$

oder auch

$$P(c < \varepsilon) = F(\varepsilon) = e^{-e^{-(\varepsilon+\alpha)}}$$

wobei α der Lageparameter und c eine Zufallsvariable der Gumbelverteilung ist. Weiterhin sei

$$F(\varepsilon_1, ..., \varepsilon_n) = \prod_{i=1}^{n} F(\varepsilon_i) = \prod_{i=1}^{n} e^{-e^{-(\varepsilon_i+\alpha_i)}}$$

mit n als Anzahl der Alternativen.
Dann gilt

$$P(U_j) = \frac{e^{\eta_j}}{\sum_{i=1}^{n} e^{\eta_i}}. \tag{5.65}$$

Beweis:
Der Beweis ist aus [MAD83] entnommen.
Gegeben sei ein Individuum mit Charakteristik s und eine Auswahl von Alternativen x aus einer Menge von Alternativen B. Es sei

$$P(x \mid s, B) \tag{5.66}$$

die Wahrscheinlichkeit dafür, dass ein Individuum mit Charakteristik s die Alternativen $x \subseteq B$ wählt. Weiterhin gelte die Aufteilung der Nutzenfunktion nach McFadden

$$U(s,x) = \eta(s,x) + \varepsilon(s,x)$$

mit ε ist unabhängig gumbelverteilt. Aus der Voraussetzung des Lemmas ist bekannt, dass $\varepsilon_i + \eta_i$ (und damit U_i) eine Gumbelverteilung mit Lageparameter $\alpha_i - \eta_i$ besitzt, wie nachfolgend aufgeführt ist

$$\begin{aligned} F_{U_i}(\varepsilon) &= P(\varepsilon_i + \eta_i < \varepsilon) \\ &= P(\varepsilon_i < \varepsilon - \eta_i) \\ &= e^{-e^{-(\varepsilon+\alpha_i-\eta_i)}}. \end{aligned}$$

Seien nun 2 Alternativen und 2 resultierende Nutzenfunktionen gegeben. Nach McFadden wird ein Individuum sich für Alternative 1 entscheiden, wenn $U_1 > U_2$. Allgemeiner ausgedrückt: Existieren n Alternativen, so wird Alternative j ausgewählt, wenn $U_j \in$

argmax$\{U_i\}_{i=1}^n$. Speziell gilt für den Fall $n = 2$

$$P(\text{Alternative } 1 \text{ wurde ausgewählt}) = P(U_1 > U_2) = P(\varepsilon_1 + \eta_1 > \varepsilon_2 + \eta_2).$$

Da ε unabhängig gumbelverteilt ist, können bezüglich der Wahrscheinlichkeit wesentlich präzisere Aussagen getroffen werden

$$\begin{aligned} P(\varepsilon_1 + \eta_1 > \varepsilon_2 + \eta_2) &= P(\varepsilon_1 + \eta_1 - \eta_2 > \varepsilon_2) \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1) \left(\int_{-\infty}^{\varepsilon_1 + \eta_1 - \eta_2} f(\varepsilon_2)\, d\varepsilon_2 \right) d\varepsilon_1 \\ &= \int_{-\infty}^{\infty} f(\varepsilon_1)\, e^{-e^{-(\varepsilon_1 + \eta_1 - \eta_2 + \alpha_2)}} d\varepsilon_1. \end{aligned} \tag{5.67}$$

Es ist zu beachten, dass $F(\varepsilon_1) = e^{-e^{-(\varepsilon_1 + \alpha_1)}}$. Daraus folgt

$$f(\varepsilon_1) = \frac{\partial F(\varepsilon_1)}{\partial \varepsilon_1} = e^{-(\varepsilon_1 + \alpha_1)} e^{-e^{-(\varepsilon_1 + \alpha_1)}}. \tag{5.68}$$

Eingesetzt in Gleichung (5.67) ergibt sich

$$\begin{aligned} P(1 \text{ wurde ausgewählt}) &= \int_{-\infty}^{\infty} e^{-(\varepsilon_1 + \alpha_1)} e^{-e^{-(\varepsilon_1 + \alpha_1)}} e^{-e^{-(\varepsilon_1 + \eta_1 - \eta_2 + \alpha_2)}} d\varepsilon_1 \\ &= e^{-\alpha_1} \int_{-\infty}^{\infty} \left(e^{-\varepsilon_1}\right) e^{\left(-e^{-\varepsilon_1}\right)\left(e^{-\alpha_1} - e^{-(\eta_1 - \eta_2 + \alpha_2)}\right)} d\varepsilon_1 \\ &= e^{-\alpha_1} \left[\frac{1}{e^{-\alpha_1} + e^{-(\eta_1 - \eta_2 + \alpha_2)}} \right] \left[e^{\left(-e^{-\varepsilon_1}\right)\left(e^{-\alpha_1} - e^{-(\eta_1 - \eta_2 + \alpha_2)}\right)} \right]_{-\infty}^{\infty} \\ &= \frac{e^{-\alpha_1}}{e^{-\alpha_1} + e^{-(\eta_1 - \eta_2 + \alpha_2)}} \\ &= \frac{e^{\eta_1 - \alpha_1}}{e^{\eta_1 - \alpha_1} + e^{\eta_2 - \alpha_2}}. \end{aligned} \tag{5.69}$$

Dieses Ergebnis kann verallgemeinert werden, da das Maximum von $(n-1)$-Alternativen stets gumbelverteilt ist. Damit kann eine zweistufige Maximierungsaussage getroffen werden

$$\begin{aligned} P(\varepsilon_1 + \eta_1 > \varepsilon_i + \eta_i, i = 1, ..., n) &= P\left(\varepsilon_1 + \eta_1 > \max_{i=2,...,n} (\varepsilon_i + \eta_i) \right) \\ &= \frac{e^{\eta_1 - \alpha_1}}{e^{\eta_1 - \alpha_1} + ... + e^{\eta_n - \alpha_n}} \\ &= \frac{e^{\widetilde{\eta}_1}}{\sum_{i=1}^{n} e^{\widetilde{\eta}_i}} \end{aligned} \tag{5.70}$$

mit $\widetilde{\eta}_j = \eta_j - \alpha_j$.

□

Umgesetzt in die bereits eingeführte Schreibweise gilt

$$\pi_{ij} = \frac{e^{\eta_{ij}}}{\sum_{j=1}^{J_i} e^{\eta_{ij}}}. \tag{5.71}$$

Formel (5.71) ist die Basisgleichung, welche das Multinomial Logit Modell definiert.

Bemerkung 5.5.2
Im Falle $J_i = 2$ kann gezeigt werden, dass die Differenz $U_{i1} - U_{i2}$ eine logistische Verteilung besitzt und das Standardmodell der logistischen Regression entsteht.

Die allgemeine Erklärung zur Bedeutung des Multinomial Logits lautet, dass sich ein Individuum i anhand *seiner Eigenschaften* x_i für eine Alternative j aus einer Menge von J Alternativen entscheidet. Dazu ist eine spezielle Bewertung der Eigenschaften des Individuums x_i für jede Alternative im Einzelnen nötig. Diese Bewertung wird ausgedrückt durch

$$\eta_{ij} = \boldsymbol{x}_i \boldsymbol{\beta}_j \tag{5.72}$$

Der Conditional Logit

In die Sprache der Monate, Wege und Routen transferiert besteht die Anwendung des Multinomial Logit Modells (5.72) darin, dass für einen Weg i mittels *einem* $\boldsymbol{x}_i$ für jede Route ein *eigenes* $\boldsymbol{\beta}_j$ berechnet werden soll. Dies würde nur dann funktionieren, wenn x_i die Eigenschaften eines Passagieres darstellen würde, der sich für eine der Routen zu entscheiden hat. In einem solchen Fall wird bei x_i von der **Charakteristik des Wählers** gesprochen. Eine solche Aufgabenstellung liefert das behandelte Problem aber *nicht*.
Stattdessen hat ein Passagier des Weges i für jede Route j dieselben Bewertungsrichtlinien $\boldsymbol{\beta}$, welche an die Eigenschaften x_{ij} einer jeden Route j angelegt werden. Es sei zu beachten, dass $\boldsymbol{\beta}$ keinen Index enthält. Es ist also *ein* $\boldsymbol{\beta}$ für *alle* möglichen Routen j des Weges i zu bestimmen. In einem solchen Fall wird bei x_i von der **Charakteristik der Wahlen** gesprochen. Damit ergibt sich eine Neuformulierung der Gleichung (5.72)

$$\eta_{ij} = \boldsymbol{x}_{ij} \boldsymbol{\beta}. \tag{5.73}$$

Es entsteht das Modell des **Conditional Logit**

$$\pi_j(\boldsymbol{X}_i) = \frac{e^{\boldsymbol{x}_{ij}\boldsymbol{\beta}}}{\sum_{k=1}^{J_i} e^{\boldsymbol{x}_{ik}\boldsymbol{\beta}}}. \tag{5.74}$$

welches von McFadden 1973/74 beschrieben wurde und bei dem an Hand der *Eigenschaften der Alternativen* x_{ij} entschieden wird.

5.5.3 Das Verfahren des Conditional Logit

Mit Einführung der Auswahlwahrscheinlichkeit des Conditional Logit endet die Herleitung desselben. Der weitere Verlauf, die Argumentation und der theoretische Hintergrund verlaufen analog zu Abschnitt 5.4.6. Es wird die Likelihood-Funktion aufgestellt, aus der mittels Logarithmierung die logarithmische Zielfunktion, die Maximum-Likelihood-Funktion, entsteht. Von dieser ist das Maximum zu bestimmten. Dafür werden die ersten beiden Ableitungen nach den Elementen von $\boldsymbol{\beta}$ gebildet, welche den Gradienten und die Hessematrix darstellen. Mittels Taylorreihenentwicklung wird über das Newton-Raphson-Verfahren das Maximum der Maximum-Likelihood-Funktion iterativ angenähert. Erneut ist die Berechnung der inversen Hessematrix nötig. Dazu wird die Iterationsvorschrift des Newton-Raphson-Verfahrens in ein Gleichungssystem umgewandelt, welches iterativ über den Newton-konjugierten Gradientenabstieg das Maximum der Taylorreihenentwicklung näherungsweise bestimmt.
Es bleibt, die Formeln derjenigen Funktionen anzugeben, die sich aufgrund der neuen Massenfunktion geändert haben.

Bemerkung 5.5.3
Die Formeln wurden nicht in Matrix- und Vektorschreibweise zusammengefasst, da die Lesbarkeit dadurch zum Teil deutlich eingeschränkt ist.

Modellfunktion für die Auswahlwahrscheinlichkeit

$$\pi_j(\boldsymbol{X}_i) = \frac{e^{\boldsymbol{x}_{ij}\boldsymbol{\beta}}}{\sum_{k=1}^{J_i} e^{\boldsymbol{x}_{ik}\boldsymbol{\beta}}}.$$

Likelihood-Funktion

$$\begin{aligned}\text{Likelihood-Funktion} &= \prod_{i=1}^{N} \frac{R_i!}{\prod_{k=1}^{J_i} y_{ik}!} \prod_{j=1}^{J_i} \pi_j \left(\boldsymbol{X}_i\right)^{y_{ij}} \\ &= \prod_{i=1}^{N} \frac{R_i!}{\prod_{k=1}^{J_i} y_{ik}!} \prod_{j=1}^{J_i} \left(\frac{e^{\boldsymbol{x}_{ij}\boldsymbol{\beta}}}{\sum_{k=1}^{J_i} e^{\boldsymbol{x}_{ik}\boldsymbol{\beta}}} \right)^{y_{ij}} . \end{aligned} \tag{5.75}$$

Da nun mehr als zwei Auswahlmöglichkeiten existieren wird die Binomialverteilung zu einer Multinomialverteilung.

Maximum-Likelihood-Funktion (logarithmische Zielfunktion)

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[G + \sum_{j=1}^{J_i} y_{ij} \left(\sum_{l=1}^{p} x_{ijl}\beta_l - \log \left(\sum_{k=1}^{J_i} e^{\sum_{l=1}^{p} x_{ikl}\beta_l} \right) \right) \right] \tag{5.76}$$

mit der Konstanten

$$G = \sum_{m=1}^{R_i} \log(m) - \left(\sum_{j=1}^{J_i} \sum_{m=1}^{y_{ij}} \log(m) \right) \tag{5.77}$$

Bemerkung 5.5.4
Zur Vermeidung von Irritationen beim Lesen der Fachliteratur sei angemerkt, dass zwei verschiedene Varianten der Maximum-Likelihood-Funktion existieren. Sie unterscheiden sich lediglich in der Verwendung von G. So wird G, welcher aus dem Multinomialkoeffizienten von Gleichung (5.75) entsteht, in älteren Skripten wie bei McFadden 1973 verwendet [FAD73]. Jüngere Skripte wie Hoffmann und Duncan 1988 [HOF88] oder Agresti 2007 [AGR07] verzichten auf die Verwendung von G, da er bereits bei der ersten Ableitung (5.78) entfällt und somit keinen Einfluss auf das Ergebnis besitzt. Der Vollständigkeit halber ist er aber mit angegeben.

Erste Ableitung der Maximum-Likelihood-Funktion (Elemente des Gradienten)

$$u_a^{(t)} = \frac{\partial L\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \beta_a^{(t)}} = \sum_{i=1}^{N} \sum_{j=1}^{J_i} y_{ij} \left(x_{ija} - \frac{\sum_{k=1}^{J_i} x_{ika} e^{\sum_{l=1}^{p} x_{ikl}\beta_l^{(t)}}}{\sum_{k=1}^{J_i} e^{\sum_{l=1}^{p} x_{ikl}\beta_l^{(t)}}} \right) . \tag{5.78}$$

Zweite Ableitung der Maximum-Likelihood-Funktion (Elemente der Hessematrix)

$$
\begin{aligned}
h_{ab}^{(t)} &= \frac{\partial^2 L\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \beta_a^{(t)} \partial \beta_b^{(t)}} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{J_i} y_{ij} \left(\frac{\left(\sum\limits_{k=1}^{J_i} x_{ika} e^{\sum\limits_{l=1}^{p} x_{ikl}\beta_l^{(t)}} \right) \left(\sum\limits_{k=1}^{J_i} x_{ikb} e^{\sum\limits_{l=1}^{p} x_{ikl}\beta_l^{(t)}} \right)}{\left(\sum\limits_{k=1}^{J_i} e^{\sum\limits_{l=1}^{p} x_{ikl}\beta_l^{(t)}} \right)^2} \right. \\
&\quad \left. - \frac{\left(\sum\limits_{k=1}^{J_i} x_{ika} x_{ikb} e^{\sum\limits_{l=1}^{p} x_{ikl}\beta_l^{(t)}} \right) \left(\sum\limits_{k=1}^{J_i} e^{\sum\limits_{l=1}^{p} x_{ikl}\beta_l^{(t)}} \right)}{\left(\sum\limits_{k=1}^{J_i} e^{\sum\limits_{l=1}^{p} x_{ikl}\beta_l^{(t)}} \right)^2} \right).
\end{aligned}
\tag{5.79}
$$

5.5.4 Unterklassen des Conditional Logit

Es sei wiederholt, dass der Conditional Logit davon ausgeht, dass die Alternativen unabhängig voneinander sind. Die genaue Bezeichnung der Unabhängigkeit lautet dabei „Independence of Irrelevant Alternatives“ (**IIA**). Laut [MAI90] und [BAS] ist dieser Fall speziell für den Multinomialen Logit und damit auch für den Conditional Logit zu beachten.
Bei Betrachtung zweier Alternativen besagt die IIA, dass die Nutzen anderer Alternativen keinen Einfluss auf die Quotienten zweier Auswahlwahrscheinlichkeiten besitzt. Diese Aussage liefert die vorausgesetzte Unabhängigkeit der stochastischen Nutzenkomponente. Bei Veränderungen der Parameter einer Route ergeben sich als Folge der IIA konstant bleibende Nutzenkomponenten der anderen Routen.
In der Anwendung hat sich gezeigt, dass sowohl der Multinomiale Logit, als auch der Conditionl Logit robust sind gegenüber Verletzungen der IIA. Da die angegebene Beschreibung leicht falsch zu verstehen ist, folgt eine weitere Erklärung nach [HAR04] welche in den Termen des bekannten „Roter-Bus-Blauer-Bus“-Phänomens wiedergegeben werden soll. Dazu stehen zwei deutlich verschiedene Transportalternativen zur Verfügung. Nun wird eine weitere Alternative angeboten, welche sich von einer der bisher existierenden Alternativen kaum unterscheidet. Diese Wahlmöglichkeiten seien der Transport per Auto, der Transport per rotem Bus und der Transport per blauem Bus. Stehen nur Auto und roter Bus als Alternativen zur Verfügung, erhält das Auto zum Beispiel eine Bewertung per Nutzenkomponente von 8 und der rote Bus eine 4. Bei einem zusätzlichen blauen Bus sollten sich die Busse nach logischem menschlichen Denken die Auswahlwahrscheinlichkeit des roten Busses teilen, da zwischen ihnen kein wesentlicher Unterschied herrscht und beide das selbe Transportfahrzeug verkörpern. Bei gültiger IIA sind sie jedoch nicht identisch, da jedes Fahrzeug eine eigene Alterna-

tive darstellt. Daher wird der rote Bus bei der Berechnung der Nutzenkomponente des blauen Busses nicht derart mit einbezogen, wie die Gesetze der menschlichen Logik es verlangen. Roter und blauer Bus erhalten somit eine Bewertung von jeweils 4. Daher ist es nach McFadden wichtig, irrelevante Alternativen für den Conditional Logit auszuschließen. Im vorliegenden Fall sollte die IIA nicht verletzt werden, da jede Route allein durch ihre Flughäfen einzigartig ist und sich von den anderen Alternativen unterscheidet. Dies ist in anderen Anwendungen nicht immer möglich. Daher existieren mehrere Weiterentwicklungen des Conditional Logits, um dieses Problem zu umgehen.
Der **Conditional Probit** geht davon aus, dass die Fehlerterme normalverteilt sind, siehe [HEC06] oder auch [HAU78]. Das Problem dabei ist, dass bei der Anpassung über Maximum-Likelihood multivariate Integrale zu lösen sind; mit einem Integral für jede Alternative. Diese Variante ist nur für Probleme mit weniger als 5 Alternativen empfohlen. Obwohl es über die Jahrzehnte verschiedene Lösungsansätze gab (unter anderem rekursive Varianten zu Berechnung der Integrale), wurde noch kein effizientes Verfahren zur Lösung dieses Problems gefunden.
Die beliebtere Variante ist deswegen der **Nested Logit**, siehe dazu [SIL] und [IMB07]. Dieser unterteilt die Alternativen in sogenannte *Nester*. Ein Nest besteht dabei aus Alternativen, die sich in einem oder mehreren Merkmalen unterscheiden, welche die Annahme der IIA verletzen. Diese speziellen Netzmerkmale werden innerhalb des Nestes auf einen fixen Wert gesetzt, sodass die IIA im Nest erfüllt ist. Die Berechnung der Auswahlwahrscheinlichkeit erfolgt dann über bedingte Wahrscheinlichkeiten in einem zweistufigen Wahrscheinlichkeitsbaum (erste Stufe Nester mit Nesteigenschaften (Eigenschaften die das Nest ausmachen und von anderen Nestern unterscheiden), zweite Stufe Alternativen in einem Nest mit Alternativeneigenschaften ohne Nesteigenschaften). Wenn i das Nest beschreibt und j die Alternative in einem Nest, so berechnet sich die Wahrscheinlichkeit für eine bestimmte Alternative j mittels $P(j) = P(\text{Nest } i) \cdot P(j|\text{Nest } i)$. Bei dieser Variante ist das Wissen darüber vorausgesetzt, welche Alternativen wegen welchen Eigenschaften zu Nestern zusammenzufassen sind. Die Berechnung der Wahrscheinlichkeit und die Anpassung erfolgen dabei analog zum Conditional Logit. Zu beachten ist dabei, dass sich $\underline{x}_{ij}$ in zwei Vektoren aufsplittet.

5.5.5 Vor- und Nachteile

Wie eingangs des Abschnittes bereits gesagt, bildet dieses Modell eine Erweiterung des Modells 4 aus dem vorherigen Abschnitt 5.4. Es gibt deswegen keine weiteren neuartigen Erkenntnisse; alle wichtigen Themen sind bereits in Abschnitt 5.4.8 ausführlich dargelegt und müssen lediglich nocheinmal erwähnt werden. Der Großteil der Nachteile der logistischen Regression ist mit diesem Modell aufgehoben. Bestehende Vorteile bleiben erhalten.

Den Hauptvorteil dieses Modells bildet, wie schon beim Grundmodell, die Einbeziehung mehrerer Eigenschaften einer Route, allerdings sind erneut Analysen durchzuführen,

um die Bedeutung der einzelnen Inputparameter zu bestimmen.
Weiterhin kann das Verfahren zur Vermeidung einer unzureichenden Anzahl von Beobachtungen zu einer allgemeinen Schätzung erweitert werden, sodass ein $\boldsymbol{\beta}_{\text{ges}}$ ermittelt wird, welches auf jeden Weg angewandt werden kann. Dieses $\boldsymbol{\beta}_{\text{ges}}$ kann als Startwert für die Berechnungen der wegspezifischen $\boldsymbol{\beta}$ dienen. Anstatt die Werte mehrerer Monate für die Berechnung eines speziellen $\boldsymbol{\beta}$ zu verwenden, wird ein einzelner Monat betrachtet. Dabei steht der i-Wert nicht mehr für verschiedene Monate, sondern für verschiedene Wege. Die Realisierungen der von i abhängigen Variablen sind diesem Kontext anzupassen. Am Vorgehen ändert sich nichts. Dies führt erneut zu einer Erhöhung der Stabilität der Lösung gegenüber Ausreißern, die Rechenzeit wird heruntergesetzt und ein aktueller Monat verwendet
Im Vergleich zur logistischen Regression gereichen die Modellvoraussetzungen nun zum Vorteil. Das Modell ist als Vertreter der multinomialen Regression auf den multiplen, nichtbinären Output ausgelegt. In der Fachliteratur wird jedoch darauf hingewiesen, dass es weiterhin anfällig für das „Roter-Bus-Blauer-Bus“-Phänomen ist.
Im Gegensatz zum vorherigen Modell müssen die Ergebnisse für einzelne Routen in der Summe nicht zu eins normiert werden. Dies geschieht automatisch. Damit liefert die Ausgabe des Modells direkt die gesuchten relativen Häufigkeiten in Form der Auswahlwahrscheinlichkeiten der einzelnen Routen.
Die in Abschnitt 5.4.8 bei Nachteil 4 bemängelte Abhängigkeit der Daten wurde in die Berechnung verschoben. Laut [AGR02] hat sich die Relevanz dieses Problems damit deutlich abgeschwächt, da die Existenz mehrerer konkurrierender Auswahlmöglichkeiten Teil des Modells ist.

Die zwei verbleibenden Nachteile betreffen die Komplexität und die Geschwindigkeit. Das Modell ist langsamer als die logistische Regression. Das liegt einerseits an den geschachtelten Iterationen wie bei Modell 5, andererseits (und das zu einem wesentlich größeren Teil) aber auch an der erhöhten Anzahl an Matrizenmultiplikationen, die aufgrund der geringeren Komplexität bei der logistischen Regression nicht durchgeführt werden mussten.
Wie in Abschnitt 4.4 erklärt, steigt die Modellkomplexität mit der verbesserten Anpassung des Modells an die Problembedingungen. Dabei besteht die Gefahr der Instabilität des Modells, wenn die wahre Input-Output-Beziehung von der Modellannahme abweicht.

6 Die Vormodelle

Mit Abschluss des letzten Kapitels wurden ausreichend Modelle zur Lösung des im Abschnitt 2.2.2 vorgestellten Primärproblems angegeben. Dabei sollen Vorhersagen zu einem Monat in der nächsten Zukunft getroffen werden. Für diesen muss vorausgesetzt werden, dass die Schedule-Daten vorliegen.
Sollte sich der zu schätzende Monat nicht in der näheren Zukunft befinden, so liegen die Schedule-Daten unter Umständen nicht vor. Andernfalls ist damit zu rechnen, dass die Schedule-Daten aufgrund von Routenplanungen von Seiten der Airlines bereits erstellt wurden. Zudem kann es sein, dass Wege oder Routen ausfallen, da bessere oder andere Routen aktiv sind, sich das Passagierinteresse verlagert hat, neue Flughäfen erbaut worden sind (zum Beispiel in China, mehr als 60 in den letzten 15 Jahren). Ein anderes Anwendungsgebiet sind Zukunftsprognosen für Airlines, die entweder zukünftige Entwicklungen bestimmen, verschiedene Szenarien oder verschiedene Flugpläne testen wollen. Dabei kennen sie ihre eigenen Flugpläne und Schedule-Daten, die der anderen Airlines sind aber logischerweise unbekannt. Diese können mittels der Vormodelle geschätzt werden.

Bemerkung 6.0.5
Bei den nachfolgenden drei Modellen ist lediglich das erste in der Lage, neben der Vorhersage der Existenz/Nichtexistenz von Routen fehlende Schedule-Daten zu schätzen. Daher wird nach der Durchführung der Modelle zwei und drei Modell eins ohne Routenexistenzbestimmung durchgeführt und die fehlenden Daten in den zuvor bestimmten Routen ergänzt.

6.1 Vormodell 1: Naiver Vergleich

Das vorliegende Modell ist das erste einer Reihe von drei Modellen, die sich mit der Vorhersage der Existenz von Routen beschäftigen. Es ist das einfachste Modell, da es keine mathematischen Verfahren benutzt außer dem direkten Vergleich von Daten und Zufallszahlen. Die grobe Idee lautet wie folgt: Der naive Vergleich bedient sich zweier Inputmonate, denen Informationen entnommen werden. Diese Informationen beinhalten Daten über den Anstieg, die Absenkung oder den Stillstand bestimmter *Segmentwerte* in Form relativer Häufigkeiten (Definition Segment in Definition 2.2.3 in Abschnitt 2.2.3). Anschließend werden die Daten des Referenzmonats bestimmt. Für jeden untersuchten Segmentwert jedes Segmentes jeder Route des Referenzmonats wird anschließend über die Ziehung einer Zufallszahl entschieden, ob und wie er sich ändert. Es kann der Fall auftreten, dass ein oder mehrere Werte nicht mehr existent sind. In diesem Fall wird die Route entfernt. Das Ergebnis ist eine Route des zu schätzenden Monats mit allen Merkmalen der DPD, welche höchstwahrscheinlich nicht identisch zu den Inputmonaten

ist.

6.1.1 Die Daten

Die Segmentwerte, die für dieses Modell relevant sind, lauten:

- Flugzeit - Angabe in Prozent, aufgeteilt in die Prozentklassen:
  „no entry“, 0 - 5%, 6 - 15%, 16 - 30%, 31 - 50%, 51 - 100%
- Frequenz - Aufgeteilt in die Frequenzklassen:
  „no entry“, 0 - 5, 6 - 15, 16 - 35, 36 - 60, $>$60
- Minimaler Flugzeugtyp
  „no entry“, 0 - 15
- Maximaler Flugzeugtyp
  „no entry“, 0 - 15

Alle anderen Informationen sind nicht von Belang, da sie topografisch gesehen statisch sind. So ändert sich weder die Entfernung zwischen zwei Flughäfen noch ist davon auszugehen, dass das Gebiet des Flughafens einem anderen Land zuzuordnen ist. Weitere routenabhängige Informationen wie die Anzahl der Zwischenstops können ebenfalls keinen Änderungen unterliegen.
Es kommt allerdings sehr wohl vor, dass auf einem bestimmten Segment verschiedene Flugzeuge eingesetzt werden, welche eine unterschiedliche Flugzeit benötigen oder von verschiedenen Airlines unterschiedlich frequentiert werden.
Die DD und DPD basierten Werte fließen nicht in die Betrachtung mit ein, da sie als Teil des Primärproblems geschätzt werden müssen beziehungsweise davon auszugehen ist, dass sie vorliegen.

6.1.2 Variablen

Neben allgemeinen Bezeichnungen seien in Tabelle 6.1 sechs Zählvariablen a, b, c, d, e, f eingeführt, davon seien a, b einfache Variablen und c, d, e, f Vektoren, die für jedes Segment einzeln gelten:

Variable	Beschreibung
A	Älterer Inputmonat
B	Jüngerer Inputmonat
C	Referenzmonat für die Daten von D
D	Der zu schätzende Monat
$a_{(\text{Segment})}$	Anzahl der Segmente, die beim Vergleich von A und B auftreten und nicht den Eintrag „no entry“ besitzen
$b_{(\text{Segment})}$	Anzahl der Segmente, die beim Vergleich von A und B nicht in B auftreten oder den Eintrag „no entry“ besitzen
$c_{(\text{Segment})}$	1×3 - Vektor, der zählt, ob sich beim Vergleich von A und B die Flugzeit des Segmentes: Nicht verändert, verringert, vergrößert hat
$d_{(\text{Segment})}$	1×3 - Vektor, der zählt, ob sich beim Vergleich von A und B die Frequenz des Segmentes: Nicht verändert, verringert, vergrößert hat
$e_{(\text{Segment})}$	1×3 - Vektor, der zählt, ob sich beim Vergleich von A und B der minimale Flugzeugtyp des Segmentes: Nicht verändert, verringert, vergrößert hat
$f_{(\text{Segment})}$	1×3 - Vektor, der zählt, ob sich beim Vergleich von A und B der maximale Flugzeugtyp des Segmentes: Nicht verändert, verringert, vergrößert hat

Tabelle 6.1: Variablen dieses Abschnittes

Bemerkung 6.1.1
Zur verbesserten Lesbarkeit sei der Index „Segment“ mit „Seg“ abgekürzt.

6.1.3 Schritt 1: Die Aufstellung der relativen Wahrscheinlichkeiten

Das Modell benötigt zwei Referenzmonate A und B, wobei A der ältere Monat ist. Die Zählvariablen a, b, c, d, e, f sind mit dem Wert 0 zu initiieren.
Diese Variablen werden verändert, indem nun jede Route von A mit derselben Route aus B verglichen wird. Dabei können folgende Fälle eintreten:

1. A-Route existiert nicht, B-Route existiert

2. A-Route existiert, B-Route existiert nicht

3. A-Route existiert, B-Route existiert

Der erste Fall ist nicht von Belang, da zeitlich voranschreitende Änderungen beobachtet werden. Es können aber keine solche Aussage über eine neu eingeführte Route getroffen werden. Oder es ist nicht unüblich, dass sie unterbrochen wird. Dann ist davon auszugehen, dass sie erneut ausfällt. So oder so können keine Informationen gewonnen werden.

Im zweiten Fall wird $a_{(\text{Seg})}$ für jedes Segment der Route um 1 erhöht, da sie aufgetreten sind. $b_{(\text{Seg})}$ wird ebenfalls für jedes Segment der Route um eins erhöht. da der Ausfall der Route mit dem Ausfall aller Segmente gleichgesetzt. Warum: In der Realität fällt eine Route meist wegen des Ausfalles eines einzelnen Segmentes aus. Es liegt aber keine Information darüber vor, welche Route dies ist. Zur Vereinfachung werden alle Segmente als ausgefallen angenommen.
$a_{(\text{Seg})}$ beinhaltet die Gesamtanzahl des Auftretens des Segmentes über alle Routen hinweg. $b_{(\text{Seg})}$ zählt die Anzahl des Ausfalls eines Segmentes. Die Variable wird benötigt, um später die relative Häufigkeit für den Ausfall des Segmentes zu bestimmen.

Der letzte Fall ist der günstigste. Die Zählvariable $a_{(\text{Seg})}$ jedes auftretenden Segmentes der Route wird um 1 erhöht. Nicht aber bei $b_{(\text{Seg})}$, da die Route nicht ausgefallen ist. Anschließend wird für jeden Wert jedes Segmentes der Route per Addition einer 1 in $c_{(\text{Seg})}, d_{(\text{Seg})}, e_{(\text{Seg})}, f_{(\text{Seg})}$ eingetragen, ob der er in B noch in derselben Klasse, eine Klasse höher oder eine Klasse tiefer liegt. Siehe dazu das Beispiel 6.1.1. Sollte der Eintrag „no entry“ lauten, ist doch ein Eintrag auf eben beschriebene Art und Weise bei $b_{(\text{Seg})}$ vorzunehmen.

Beispiel 6.1.1
Die Route TGL – FRA – CDG – LHR aus Beispiel 2.2.1 auf Seite 5 wird behandelt. Es sei angenommen, dass die Route sowohl in A als auch in B existiert. Dann gelten folgende Schritte. Zur Vermeidung unnötiger Wiederholungen wird lediglich das Vorgehen für das Segment (TGL, FRA) betrachtet.

$$a_{\textit{(TGL, FRA)}} \rightarrow a_{\textit{(TGL, FRA)}} + 1$$

Würde die Route nicht in B existieren, gilt zusätzlich:

$$b_{\textit{(TGL, FRA)}} \rightarrow b_{\textit{(TGL, FRA)}} + 1$$

Dafür wäre allerdings der folgende Teil nicht mehr durchzuführen:

Flugzeit	Frequenz	min. Flugzeugtyp	max. Flugzeugtyp
0.24	13	6	12

Tabelle 6.2: Werte des Segmentes (TGL, FRA) für Referenzmonat A

Flugzeit	Frequenz	min. Flugzeugtyp	max. Flugzeugtyp
0.24	12	3	13

Tabelle 6.3: Werte des Segmentes (TGL, FRA) für Referenzmonat B

Der erste Eintrag hat sich nicht verändert, es ist eine Eintragung im ersten Element des Flugzeit-Vektors vorzunehmen. Der zweite Eintrag hat sich zwar verändert, liegt aber immer noch in derselben Klasse. Dieser Fall wird damit wie der erste Eintrag behandelt. Der dritte Eintrag ist in B kleiner als in A. Es ist also eine Eintragung im zweiten Element des minimalen Flugzeugtyps vorzunehmen. Der vierte Eintrag ist in B größer als in A. Es ist also eine Eintragung im dritten Element des maximalen Flugzeugtyps vorzunehmen:

$$
\begin{aligned}
c_{\text{(TGL, FRA)}} &\rightarrow c_{\text{(TGL, FRA)}} + (1,0,0) \\
d_{\text{(TGL, FRA)}} &\rightarrow d_{\text{(TGL, FRA)}} + (1,0,0) \\
e_{\text{(TGL, FRA)}} &\rightarrow e_{\text{(TGL, FRA)}} + (0,1,0) \\
f_{\text{(TGL, FRA)}} &\rightarrow f_{\text{(TGL, FRA)}} + (0,0,1).
\end{aligned}
$$

Abschließend ist $b_{\text{(Seg)}}$ durch $\frac{b_{\text{(Seg)}}}{a_{\text{(Seg)}}}$ und die Zählvektoren durch $\frac{\text{Zählvektor}}{\sum_{i=1}^{3} \text{Zählvektor}_i}$ zu ersetzen, um die jeweiligen relativen Häufigkeiten für den Ausfall beziehungsweise den Wertebestand, -verfall oder -wachstum eines Wertes zu bestimmen.

6.1.4 Schritt 2: Die Anwendung der relativen Wahrscheinlichkeiten auf C

Nach der Beendigung der Vorarbeit können nun die Routen und ihre Werte für den zukünftigen zu schätzenden Monat D bestimmt werden. Dazu sind die Routen von C heranzuziehen und die Zählvariablen auf die Werte jedes einzelnen Segmentes anzuwenden. Alternativ kann an Stelle des Referenzmonats C erneut B verwendet werden.

Das Vorgehen lautet:

1. Wähle die Routen eines bestimmten Weges

2. Verwirf alle Routen, die wenigstens ein Segment besitzen, welches in Schritt eins nie aufgetreten ist.

3. Wiederhole das folgende Vorgehen, bis wenigstens eine gültige Route zustande kommt:

    a) Wähle eine Route, die in diesem Durchlauf noch nicht behandelt wurde

    b) Verfahre nacheinander mit jedem Segment wie beschrieben:

c) Ziehe eine Zufallszahl aus $[0,1]$. Ist der Wert kleiner als $\frac{1}{\text{Segmentanzahl}} \cdot b_{(\text{Seg})}$, so fällt das Segment aus und die Route ist ungültig. $\frac{1}{\text{Segmentanzahl}}$ deshalb, da dieser Vorgang Segmentanzahl-oft wiederholt wird und die Ausfallwahrscheinlichkeit damit Segmentanzahl-mal höher liegt.

d) Bestimme die Werte des Segmentes. Da das Vorgehen für jeden Wert analog verläuft, sei lediglich $c_{(\text{Seg})}$ beschrieben. Ziehe dazu eine Zufallszahl aus $[0,1]$. Fällt diese Zahl in den Bereich $\left[0, c_{(\text{Seg})1}\right]$, so wird der Wert aus B übernommen. Ist es der Bereich $\left(c_{(\text{Seg})1}, c_{(\text{Seg})1} + c_{(\text{Seg})2}\right]$, muss der Mittelwert der nächsthöheren Klasse des Wertes von B eingetragen werden. Analog die nächsttiefere Klasse beim Bereich $\left(c_{(\text{Seg})1} + c_{(\text{Seg})2}, 1\right]$. Ist der Wert bereits an der oberen oder unteren Ende der Klassenskala, so wird er nicht verändert. „no entry“ zählt dabei als niedrigste Klasse. Es soll kein Abstieg in diese Klasse erfolgen. Stattdessen wird die nächsthöhere Klasse beibehalten. Werte die bereits bei „no entry“ sind und dort verbleiben, behalten diesen Eintrag. Sie sind scheinbar dazu „bestimmt“ oder einfach nicht wichtig genug, um in irgendeiner Datenbank erwähnt zu werden.

e) Fällt kein Segment aus und ist die Route dadurch intakt, so sind die restlichen statischen Routenwerte zu bestimmen und die Route in den zu schätzenden Monat D einzutragen.

Bemerkung 6.1.2
Die Wahl der Input- und Referenzmonate liefert einen gewissen Handlungsspielraum. Allerdings soll an dieser Stelle die Empfehlung ausgesprochen werden, entweder zwei aufeinanderfolgende Monate als Inputmonate zu wählen, oder zweimal denselben Monat mit maximal einem Jahr Abstand. Beim Referenzmonat sollte es sich um denselben Monat eines anderen Jahres oder den Vormonat handeln. Diese Wahl besitzt die meiste Aussagekraft.

6.1.5 Vor- und Nachteile

Der größte Vorteil dieses Modells ist die Möglichkeit zur Bestimmung der vier Routenwerte. Es ist aufgrund seiner Einfachheit zudem das schnellste der vorgestellten Vormodelle. Zur Erhöhung der Geschwindigkeit wurde mit den absoluten Häufigkeiten (zum Beispiel $c_{\text{Seg}1}$) gearbeitet, statt mit relativen Häufigkeiten.

Als nachteilig anzumerken ist die starke Abhängigkeit der Güte von der Wahl der Input- und Referenzmonate. Das Verfahren ist bezüglich dieser Wahl als instabil einzuordnen. Globale wie auch nationale Großereignisse, welche in den beobachteten Monaten stattgefunden haben und starke Auswirkungen auf den Flugverkehr hatten, besitzen einen großen Einfluss auf die quantitativen Aussagen des Verfahrens.

6.2 Vormodell 2: Logistische Regression

Dieses Modell verwendet die logistische Regression, um einen Entscheidungsvektor $\boldsymbol{\beta}$ zu erzeugen. Dafür sind zwei Inputmonate notwendig. Das Modell beziehungsweise $\boldsymbol{\beta}$ wird anschließend auf einen gegebenen dritten Monat angewandt, um für seine Routen zu bestimmen, ob sie im zu schätzenden Monat auftreten oder nicht. Im Gegensatz zum Vormodell des naiven Vergleichs werden keine Routenwerte erzeugt, sondern lediglich die bereits vorhandenen Werte des dritten Monats übernommen.

Die logistische Regression wird aufgrund der Problemstellung verwendet. Für jede Route ist bezüglich seiner Charakteristika die binäre Entscheidung zu treffen, ob sie im zu schätzenden Monat auftritt oder nicht. Es existieren keine multiplen Outputs oder Wahlmöglichkeiten und für jede Beobachtung (Route) ist einzeln zu entscheiden. Es ist das ideale Anwendungsgebiet für die logistische Regression.

6.2.1 Variablen

Zur Vereinfachung der Lesbarkeit des Textes werden an dieser Stelle in Tabelle 6.4 einige Variablen und Bezeichnungen eingeführt.

Variable	Beschreibung
A	Älterer Inputmonat
B	Jüngerer Inputmonat
C	Referenzmonat für die Daten von D
D	Der zu schätzende Monat
$\boldsymbol{\beta}$	$1 \times p$ - Entscheidungsvektor aus der logistische Regression
p	Anzahl der Routeninformationen
J	Anzahl aller Routen
$i\ (i = 1, ...J)$	Indexvariable, die für eine der J Routen steht
$\boldsymbol{x}_i$	$1 \times p$ - Inputvektor, enthält Routeninformationen der Route i; Nach Abschnitt 3.2.2
$\boldsymbol{X}$	$J \times p$ - Inputmatrix, gebildet aus allen $\boldsymbol{x}_i$
y_i	Binärer Output für den Input $\boldsymbol{x}_i$, 0: Route wird nicht geflogen, 1: Route wird geflogen
$\boldsymbol{y}$	$1 \times J$ - Outputvektor, gebildet aus allen y_i
a	Entscheidungswert für Existenz oder Nichtexistenz einer Route

Tabelle 6.4: Variablen dieses Abschnittes

6.2.2 Daten

Das Verfahren der logistischen Regression benötigt Inputdaten. Diese werden aus A, B und C gewonnen, wie in Abschnitt 3.2.2 beschrieben. Im Gegensatz zu dem Modell der logistischen Regression aus Abschnitt 5.4 gehen *alle* gewonnenen Informationen in die Parameterschätzung ein. Vorher wurden die Informationen x_{a1}, x_{a2} und x_{a3} nicht aufgenommen, da y_i aus ihnen bestimmt wurde. Dies ist nun nicht mehr der Fall. Zudem ist vorausgesetzt, dass alle verwendeten Monate bereits bekannt sind. Der Input $\boldsymbol{x}_i$ kann für A, B und C erstellt werden.
y_i ist lediglich im Existenzvergleich der Route i bezüglich A und B zu bestimmen. Dabei wird per Binärkodierung festgehalten, ob Route i des Monats A auch im Monat B geflogen wurde (kodiert mit 1) oder nicht (kodiert mit 0).
Bei der Informationserstellung werden die Inputdaten aller Routen zu einer Matrix $\boldsymbol{X}$ und die Outputdaten aller Routen zu einem Outputvektor $\boldsymbol{y}$ zusammengefasst.

6.2.3 Schritt 1: Parameterschätzung per logistischer Regression

Mittels der logistischen Regression soll ein Entscheidungsvektor $\boldsymbol{\beta}$ geschätzt werden, welcher für den Input $\boldsymbol{x}_i$ einer Route vorhersagt, ob die Route in Zukunft existiert oder nicht. Das Schätzverfahren erfolgt dabei analog zu Abschnitt 5.4 mit erweiterter Inputmatrix $\boldsymbol{X}$ und verändertem, binärem Outputvektor $\boldsymbol{y}$. $\boldsymbol{X}$ wird für den Monat A erstellt. $\boldsymbol{y}$ ist so zu erzeugen, wie es in Abschnitt 6.2.2 beschrieben steht.
Die Einteilung der Routen in verschiedene Wege findet nicht statt. Es wird ein $\boldsymbol{\beta}$ für alle Routen geschätzt. Die Durchführung der Parameterschätzung per logistischer Regression geschieht also nicht für jeden Weg, sondern nur ein einziges Mal für alle Routen.
Das restliche Vorgehen zur Schätzung von $\boldsymbol{\beta}$ erfolgt analog zu Abschnitt 5.4.

6.2.4 Schritt 2: Bestimmung eines Entscheidungswertes a

Nach der Schätzung von $\boldsymbol{\beta}$ wird der Vektor genutzt, um mittels Formel (5.20) $\pi(\boldsymbol{x}_i)$ für die Route i von Monat C zu bestimmen. Welcher Wert von π steht nun aber für den Ausfall oder die Intaktheit einer Route?
Dazu sei ein Entscheidungswert a bestimmt. Wenn $\pi < a$, so ist der Route eine 0 zuzuordnen, welche für ihren Ausfall steht. Im anderen Fall ist die Route intakt und sie erhält eine 1.
Zur Ermittlung eines optimalen Wertes a sind die Routeninformationen $\boldsymbol{X}$ von A und die Outputinformationen $\boldsymbol{y}$ aus Abschnitt 6.2.2 erneut verwendet. Anschließend erfolgt unten beschriebenes Vorgehen:

- Berechne π für jede Route
- Ordne a nacheinander Werte aus $[0, 1]$ zu. Beginne bei 0.01 und erhöhe schritt-

weise um 0.01. Für jedes a geschehe folgendes:

1. Ordne der Route i eine 1 zu, wenn $\pi \geq a$, sonst 0
2. Bestimme den Quotienten b aus $b = \frac{\text{Abweichungen von } \boldsymbol{y}}{\text{Übereinstimmungen mit } \boldsymbol{y}}$.

- Derjenige Wert von a, bei der b minimal wird, ist als optimale Einstellung zu wählen.

6.2.5 Schritt 3: Anwendung von $\boldsymbol{\beta}$ und a

Nach der Schätzung von $\boldsymbol{\beta}$ und a erfolgt die Anwendung der Parameter auf die Routen von C. Dafür sind die Routeninformationen von C zu bestimmen. Alternativ kann B an Stelle eines neuen Monats C verwendet werden. Anschließend wird für jede Route i $\pi(\boldsymbol{x}_i)$ berechnet und mit a verglichen. Gilt $\pi(\boldsymbol{x}_i) \geq a$, so ist die Route samt ihren Informationen als intakte Route für den Monat D einzutragen. Im gegenteiligen Fall wird die Route nicht eingetragen.

6.2.6 Vor- und Nachteile

Der Hauptvorteil dieses Verfahrens findet sich in der Einbeziehung mehrerer Eigenschaften einer Route, sodass ein kausaler Zusammenhang zwischen Einflussfaktoren und der Existenz einer Route erstellt werden kann. Bisher gab es lediglich die Möglichkeit, relative Häufigkeiten zu verwenden. Mit diesem Modell wird wichtigen Eigenschaftsparametern die Möglichkeit geboten, in die Bewertung eingehen zu können. Welche Parameter wichtig sind und welche nicht, kann durch eine anschließende Analyse der $\boldsymbol{\beta}$-Werte bestimmt werden. Das Modell ist entsprechend erweiterbar.
Im Gegensatz zum Hauptmodell ist die logistischen Regression hier optimal für die Bearbeitung der Problemstellung geeignet, da die Anwendungsvoraussetzungen diesmal erfüllt sind.

Zu den Nachteilen lässt sich sagen, dass wegen der Komplexität des Modells $\boldsymbol{\beta}$ nur bedingt für eine direkte Interpretation der Bedeutung der Einflussfaktoren geeignet ist. Wie oben angesprochen sind weitere Analysemodelle von Nöten, um die Relevanz der einzelnen Inputparameter zu bestimmen. Aussagen über direkte Korrelation und Bedeutung der Größe der Werte lassen sich nur schlecht treffen.
Ein Nachteil dieses Verfahrens ist seine Geschwindigkeit die aufgrund der zwei geschachtelten Iterationen geringer ausfällt als bei dem vorherigen Modell.
Weiterhin gibt es kein Zusammenspiel mehrerer Routen bei der Ermittlung der Auswahlwahrscheinlichkeit.

Die Aufgabe der logistischen Regression besteht in der Entscheidung der Existenz oder Nichtexistenz einer Route bei Eingabe ihrer Parameter. Vormodell 1 muss anschließend

durchgeführt werden, um die topografisch unabhängigen Daten zu ermitteln, insofern sie erforderlich sind.

6.3 Vormodell 3: Resilient-Backpropagation im Neuronalen Netz

Dieser Abschnitt beschäftigt sich mit einer komplett anderen Art des überwachten Lernens als bisher: Den sogenannten **künstlichen Neuronalen Netzen**. Dabei handelt es sich um ein mathematisches Modell, mit dem versucht wird, die neuronalen Netzstrukturen im Gehirn nachzubilden und mittels eines Lernmodells das menschliche Lernen zu simulieren. Diese Netze existieren auch für unüberwachtes Lernen. Die Nachbildung der Netzstrukturen erfolgt in einem sehr vereinfachten Rahmen, da es kein Modell der Welt vermag, die komplexen Vorgänge des Denkens und Lernens im Gehirn exakt darzustellen. Eine genaue biologische Nachbildung ist mitunter sogar hinderlich.
Angewandt wird der **Resilient-Backpropagation-Algorithmus** von M. Riedmiller und H. Braun aus dem Jahre 1993, welcher eine Weiterentwicklung des bekannten **Backpropagation-Algorithmus** ist, bei dem einige Vorgänge bezüglich der Parameterwahl automatisiert worden sind. Die Grundlagen des Backpropagation-Algorithmus wurden 1960 durch Henry J. Kelley und 1961 durch Arthur E. Bryson gelegt. Er ist wiederum eine Verallgemeinerung der sogenannten **Delta-Regel** auf mehrschichtige Netze. Dabei wird das Netz mit Inputdaten gespeist, welche einen Output erzeugen. Durch den Vergleich des erzeugten mit dem gewünschten Output entstehen Fehlerwerte. Diese werden rückwärts, also vom Ende des Netzes beginnend, eingegeben und eine Parameterkorrektur durchgeführt. Daher der Name „Backpropagation of Error“ („Fehlerrückführung“). Der Input wird analog zu Vormodell 2 aus Abschnitt 6.2 gebildet, der Output liegt im Intervall $[-1, 1]$ entspricht aber einem binären Output. Die Grundlage für all dies schuf der amerikanische Psychologe und Informatiker Frank Rosenblatt 1958 mit seinem **Perzeptron**, der einfachsten Variante eines Neuronalen Netzes.
Versucht wird, eine möglichst knappe Variante der Implementation mit allen benötigten Definitionen, Erklärungen und Herleitungen zu geben, damit der Umfang dieses Abschnittes dem Rahmen eines in der Priorität untergeordneten Themas gerecht werden kann. Für eine kurze Einführung sei die Hausarbeit von [SCH06] empfohlen. Für einen ausführlicheren, sehr anschaulichen Einstieg in das Thema sei ein Blick in [KRI05] angeraten. Der Abschnitt ist an eben jenen Artikel angelehnt. Die Abbildungen, Herleitung und Argumentation entstammt diesen beiden Skripten.

6.3.1 Biologischer Hintergrund

Das menschliche wie auch das tierische Nervensystem besteht aus Nervenzellen, den sogenannten Neuronen. Ihr Zweck besteht im Empfang und der Weiterleitung von Signalen untereinander.

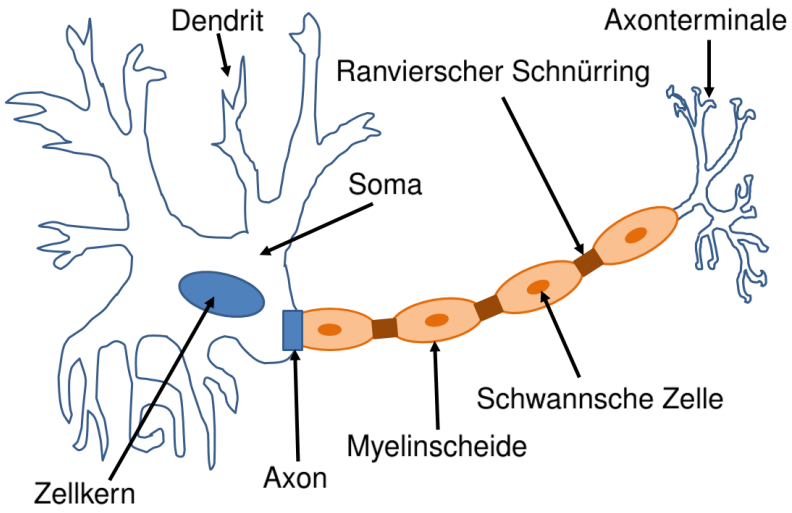


Abbildung 6.1: Schematischer Aufbau eines Neurons

In Abbildung 6.1 ist der schematische Aufbau eines Neurons zu sehen: Die Hauptbestandteile bilden der Dendrit, der Zellkörper (Soma) und das Axon. Mittels Synapsen (Axonterminale) erfolgt die Kommunikation mit anderen Neuronen. Die *Aufnahme der Informationen* anderer Neuronen findet im Dendriten statt. Am Soma findet die *Kumulation der Änderungen des Membranpotenzials* statt. Bei genügend großer Änderungsrate wird eine sogenannte *Feuerschwelle* überschritten und es erfolgt am Axonhügel die *Auslösung eines Aktionspotentials*, bei dem die gesamten Informationen weitergeleitet werden (Alles-oder-Nichts-Prinzip). Die Weiterleitung beginnt beim Axon und führt bis zu den Axonterminalen, den Endknöpfchen. Sie bilden den präsynaptischen Teil der Synapse, der Verbindung des Endknöpfchens und des Dendriten zweier Nervenzellen, wie in Abbildung 6.2 dargestellt ist.

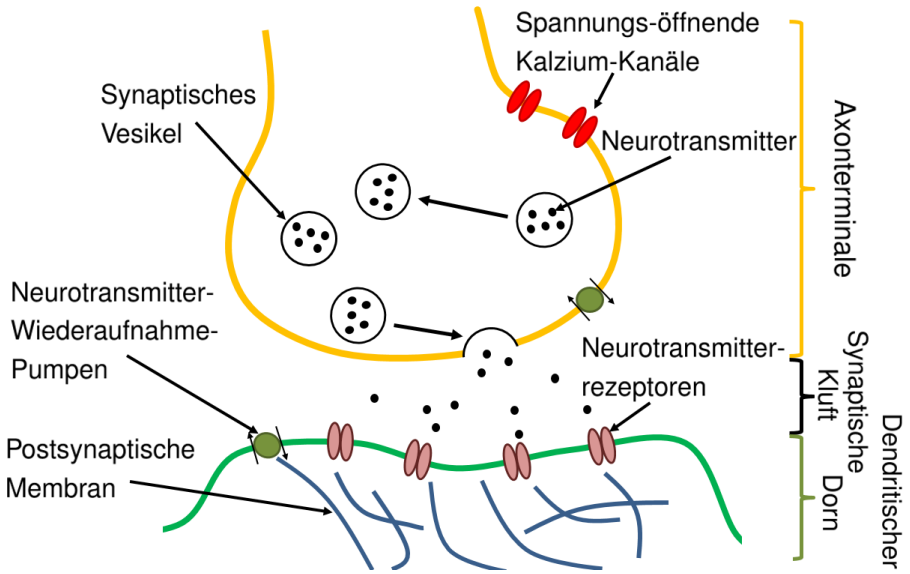


Abbildung 6.2: Schematische Darstellung der Informationsübergabe in einer Synapse

Abbildung 6.2 zeigt die *Informationsübertragung der Erregung in einer Synapse*, welche die chemische Verbindung zweier Nervenzellen darstellt. Diese chemische Übertragung geschieht mittels Ausschüttung von Neurotransmittern (zum Beispiel Acetylcholin) in den synaptischen Spalt. Diese binden sich an die Rezeptoren auf der postsynaptischen Membran des Dendriten der anderen Nervenzelle (zum Beispiel Acetylcholinrezeptoren). Die Membran reagiert mit der *Öffnung von Ionenkanälen*. Dieser Vorgang wiederum führt zur Entstehung eines elektrischen Signals, welches als Änderung des Membranpotentials an den Zellkörper weitergeleitet wird. Damit schließt sich der Kreis der Informationsweitergabe zwischen Neuronen.

Dieser biologische Vorgang, sowie der Verbund der Neuronen wird in den nächsten Abschnitten umgesetzt. Von besonderem Interesse sind dabei die kursiv dargestellten Abschnitte.

6.3.2 Das Neuronale Netz

Dieser Abschnitt beschäftigt sich mit den wichtigsten Definitionen zum Aufbau des verwendeten mathematischen Neuronalen Netzes. Einige der Definitionen können aufgrund ihrer speziellen Anpassung für das vorliegende Modell von der Fachliteratur abweichen. Zur Verbesserung der Lesbarkeit sei mit *Netz* oder *Neuronalem Netz* das mathematische Modell des künstlichen Neuronalen Netzes gemeint und nicht das biologische.

Definition 6.3.1 (Neuronales Netz)
Unter einem Neuronalen Netz ist ein sortiertes Tripel (N,V,w) *aus zwei Mengen* N *und* V *und einer Funktion* w *zu verstehen. Im graphentheoretischen Sinn ist* $N = \{i \in \mathbb{N} |$ *mit* $N \leq \mathbb{N}$ *endlich*$\}$ *eine aus den* ***Neuronen*** *gebildete Knotenmenge und zudem ist* $V = \{(i,j) | i,j \in N$ *mit* $w((i,j)) \in \mathbb{R}^+\}$ *eine gerichtete, gewichtete Kantenmenge. Die Knoten/Neuronen seien nummeriert. Das Gewicht* $w((i,j))$ *einer Kante* (i,j) *wird durch die Gewichtsfunktion* $w : V \to \mathbb{R}$ *bestimmt. Es sei mit* $w_{i,j}$ *abgekürzt.*

Bemerkung 6.3.1
An dieser Stelle sei bereits festgelegt, dass für eine Kante (i,j) $i < j$ gelten soll, um eine Ordnung der Knoten zu erhalten.

Definition 6.3.2 (Gewichtsmatrix)
Die Kantenmenge und die Gewichtungen lassen sich gleichzeitig in einer $|N| \times |N|$*-Matrix* W*, der* ***Gewichtsmatrix*** *beschreiben. Das Element an der Stelle* i,j *bezeichnet die Kante/Verbindung des Neurons* i *mit dem Neuron* j*. Durch die Orientierung der Kanten sind* (i,j) *und* (j,i) *verschieden. Aufgrund der Ordnung der Kanten aus Bemerkung 6.3.1 entsteht eine obere Dreiecksmatrix. Die Werte der Elemente geben die Stärke der Verbindung an, das* ***Kantengewicht****. Kanten mit dem Gewicht 0 sind als nicht existent anzusehen.*

Definition 6.3.3 (Multilayer-Perzeptron)
*Das **Multilayer-Perzeptron** ist eine spezielle Variante des **Feed-Forward-Netzes**, bei der die Neuronen in $n \in \mathbb{N}$ Schichten sortiert sind. Ein Feed-Forward-Netz ist ein Neuronales Netz, bei dem die Informationsübermittlung wie in Abschnitt 6.3.3 erfolgt. Eine **Schicht** („Layer“) besteht aus einer Menge von Neuronen, welche untereinander nicht verbunden sind und ohne deren Existenz das Netz in zwei Teile zerfallen würde. Das ordnungsmäßig größte Neuron einer Schicht soll kleiner als das ordnungsmäßig kleinste Neuron einer Schicht mit größeren Neuronen sein. Durch die Ordnung der Neuronen unterliegen die Schichten ebenfalls einer Ordnung. Die Neuronen zweier „nebeneinanderliegender“ Schichten sind allesamt verbunden. Dieser Zustand heißt **Vollverknüpfung**. Vor der ersten Schicht existiert ein einzelnes Neuron, das **Inputneuron** I (I für „Input“), welches mit der ersten Schicht vollverknüpft ist. Die letzte Schicht bildet eine Vollverknüpfung mit einem weiteren einzelnen Neuron, dem **Outputneuron** O (O für „Output“). Diese Neuronen bilden im graphentheoretischen Sinn eine **Quelle** und eine **Senke**. Es können mehrere Input- und Outputneuronen existieren.*

Bemerkung 6.3.2
In der Fachliteratur existieren zwei Arten von Schichten. Die erste Art beschriebt wie hier eine Menge von Neuronen. Die zweite Variante bezeichnet die Kanten zwischen zwei Neuronenmengen als Schicht. Beim Vergleich mit der Fachliteratur ist darauf unbedingt zu achten. Es ergeben sich dadurch andere Funktionen für den Backpropagation-Algorithmus.

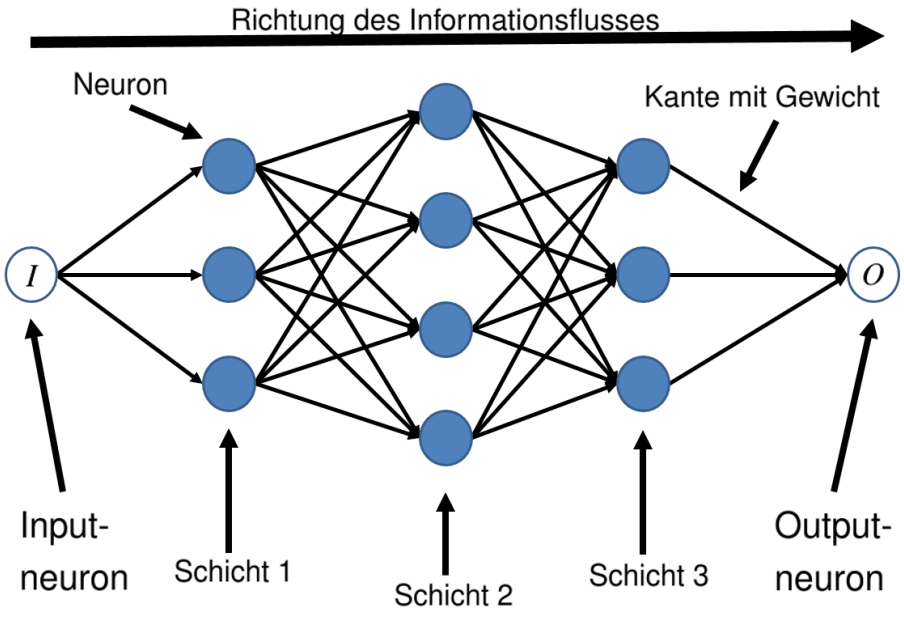


Abbildung 6.3: Schematische Darstellung eines Multilayer-Perzeptrons mit drei Schichten

6.3.3 Der Informationsfluss im Neuronalen Netz

Der Begriff Feedforward-Netz beschreibt die Art der Weitergabe von Informationen in einem Netz. Beginnend bei p **Inputneuronen** werden Informationen über gewichtete Kanten zur ersten Schicht übertragen. In dieser Schicht wird ihre Wichtigkeit für jedes einzelne Neuron bemessen. Danach erfolgt die Weitergabe in veränderter Weise an die Neuronen der nächsten Schicht. Die Informationen fließen schrittweise durch das gesamte Netz bis sie im letzten Schritt im Outputneuron münden. Der danach ermittelte Outputwert bildet den zu interpretierenden Output. Diese Art des Informationsflusses nennt sich **topologische Aktivierungsordnung**. Andere Varianten sind denkbar, für das vorgestellte Verfahren aber nicht praktikabel.
Die Inputneuronen besitzt hierbei lediglich die repräsentative Aufgabe der Initialisierung der Inputinformationen zur Eingabe in das Netz. Die Anzahl der Neuronen in dieser Schicht entspricht der Anzahl p der Inputparameter. In diesem Fall sind es alle Parameter $x_{a1} - x_{ff}$ aus Abschnitt 3.2.2. Der Outputwert eines Inputneurons entspricht also dem Wert des repräsentierten Inputparameters einer Route. Die Anzahl der Neuronen zu jeder der drei inneren Neuronenschichten beträgt ebenfalls p. Der Output bildet eine Zahl im Intervall $[-1, 1]$, die mittels Schwellwert zu einem binären Output umgewandelt wird. -1 steht dabei für den Ausfall einer Route bei Eingabe ihrer Routenparameter und 1 für die Existenz der Route. Grundsätzlich können auch mehrere Outputneuronen modelliert werden.

Bemerkung 6.3.3
Wichtig: Der in diesem Abschnitt beschriebene Vorgang bildet das Modell eines fertig trainierten Neuronalen Netzes. Bei der Anwendung werden die Inputparameter der Route eines Referenzmonats in das Netz eingegeben, indem die Werte der Inputparameter durch eine gewichtete Kante von den entsprechenden Inputneuronen zu den Neuronen der ersten Schicht repräsentiert werden. Am Ende gibt das Netz einen Wert des Intervalls $[-1, 1]$ aus. Ist dieser Wert ≥ 0, so wird die Route im zukünftigen Monat aktiv sein, ansonsten nicht. Die Anwendung eines Entscheidungswertes wie in Abschnitt 6.2 ist allerdings denkbar.

Bis auf das Inputneuron ist die Funktionsweise bei allen Neuronen identisch. Sie besteht im Allgemeinen aus der Verarbeitung der in das Neuron eingehenden Informationen in drei Schritten:

1. Mittels der **Propagierungsfunktion** werden die Outputs der Neuronen der Vorgängerschicht zu einer einzigen Information, der **Netzeingabe**/**Netzinput**, zusammengefasst.

2. Die **Aktivierungsfunktion** bewertet die Netzeingabe. Tritt der Fall ein, dass ein

bestimmter **Schwellwert** überschritten wird, so ist die Information wichtig genug, um weitergegeben zu werden.

3. Zur Weitergabe der Information an die Neuronen der Folgeschicht wird der durch die Aktivierungsfunktion ermittelte Wert an die **Outputfunktion** übergeben. Das Ergebnis bildet den Output des Neurons.

Bemerkung 6.3.4
Der vollständige Algorithmus wird am Ende in Abschnitt 6.3.8 angegeben.

Definition 6.3.4 (Propagierungsfunktion)
Sei j ein Neuron und sei $M = \{i_1, ..., i_m\}$ die Menge der Neuronen der Vorgängerschicht mit der Eigenschaft, dass $\forall z \in \{1, ..., m\} : \exists w_{i_z j}$. Sei $o_z \in \mathbb{R}$ der Outputwert eines Vorgängerneurons z. Dann berechnet die Propagierungsfunktion f_{prop} die Netzeingabe net_j von j

$$net_j = f_{prop}\left(o_{i_1}, ..., o_{i_m}, w_{i_1 j}, ..., w_{i_m j}\right). \tag{6.1}$$

In diesem Modell sei f_{prop} die **gewichtete Summe**

$$\text{net}_j = \sum_{i \in M} o_i w_{i,j}, \tag{6.2}$$

wobei $o_i w_{ij}$ der Stärke der Information entspricht, wenn sie von Neuron i an Neuron j übergeben wird. Die gewichtete Summe ist die meistverwendete Propagierungsfunktion.

Definition 6.3.5 (Aktivierungszustand)
Sei j ein Neuron. Dann bezeichnet der Aktivierungszustand a_j von j den Aktivitätsgrad des Neurons. Der Wert von a_j ist durch die Aktivierungsfunktion zu bestimmen.

Definition 6.3.6 (Schwellwert)
Sei j ein Neuron. Dann beschreibt der Schwellwert des Neurons Θ_j die Stelle der größten Steigung der Aktivierungsfunktion.

Definition 6.3.7 (Aktivierungsfunktion)
Sei j ein Neuron. Die Aktivierungsfunktion f_{act} bestimmt einen neuen Aktivierungszustand $a_{j\ neu}$ durch eine Berechnung aus der Netzeingabe net_j, dem alten Aktivierungszustand $a_{j\ alt}$ und dem Schwellwert Θ_j

$$a_{j\ neu} = f_{act}\left(net_j, a_{j\ alt} \cdot \Theta_j\right). \tag{6.3}$$

Für die Aktivierungsfunktion wurde in den frühen Entwicklungsphasen der Neuronalen Netze nach dem biologischen „Alles oder Nichts“-Vorbild die Heaviside-Funktion ver-

wendet. Da zur Anpassung jedoch die Ableitung der Aktivierungsfunktion nötig ist, trat die Logistische Funktion an die Stelle der Heaviside Funktion. D. Anguita, G. Parodi und R. Zumino schlugen 1993 die Funktion des Tanges hyperbolicus vor. Er ermöglichte eine Berechnung mittels Computer in der 200-fachen Geschwindigkeit im Vergleich zur logistischen Funktion. Sein Wertebereich liegt zwar „nur“ im Intervall $(-1,1)$ statt in $[-1,1]$. Dafür ist die Funktion der ersten Ableitung mit $\tanh(x)' = 1-(\tanh(x))^2$ simpel und schnell zu berechnen.

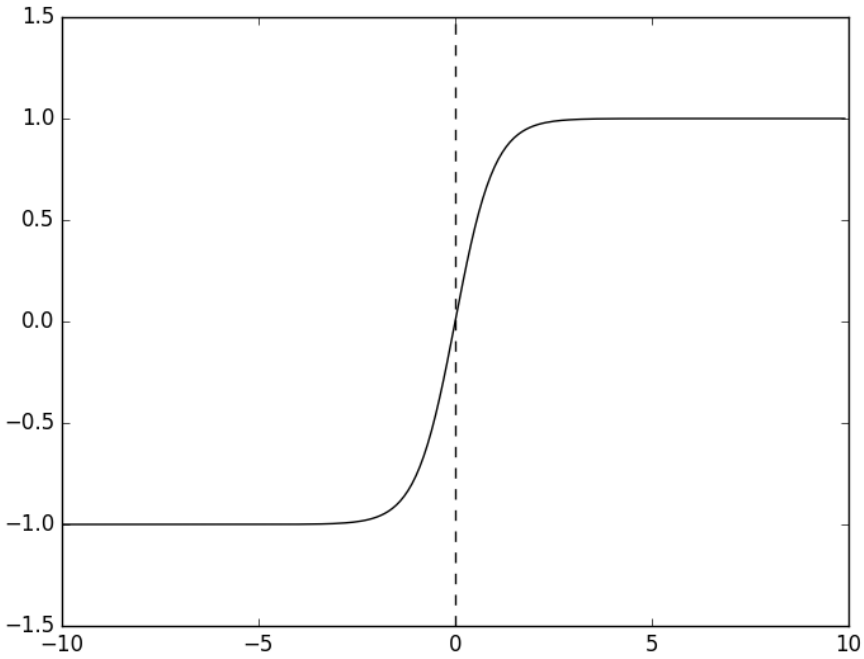


Abbildung 6.4: Der Graph des Tangens hyperbolicus

Der alte Aktivierungszustand $a_{j\,\text{alt}}$ wird in diesem Modell nicht benötigt und ist somit kein Teil des Inputs. Durch die Erstellung eines **Schwellwertneurons** oder auch **Biasneuron** wird Θ_j Teil von net_j. Daher entfällt Θ_j ebenfalls aus dem Input. Somit ist der Aktivierungszustand lediglich von der Netzeingabe net_j abhängig. Es gilt

$$a_j = \tanh(\text{net}_j). \tag{6.4}$$

Erklärung zum Schwellwertneuron:
Für viele Anwendungen ist es sehr kompliziert, zusätzlich zu den Kantengewichten die Schwellwerte anzupassen. Deshalb wird das Modell von Abbildung 6.3 um ein Neuron in der Inputschicht erweitert. Dieses Schwellwertneuron besitzt eine Verbindung zu *jedem* Neuron aller Folgeschichten, inklusive der Outputschicht. Sein Output ist stets 1

und das Kantengewicht zu einem Neuron j wird mit $-\Theta_j$ initiiert. Die Anpassung der Schwellwerte erfolgt damit automatisch durch die Anpassung der Kantengewichte.
Das Schema des veränderten Modells sieht dadurch unübersichtlich aus. Daher wird das Schwellwertneuron oftmals weggelassen, im Wissen, dass es eigentlich modelliert ist.

Definition 6.3.8 (Outputfunktion)
Sei j ein Neuron. Die Outputfunktion f_{out} berechnet aus dem Aktivierungszustand a_j den Outputwert o_j, der an die Neuronen der Nachfolgeschicht weitergegeben wird

$$o_j = f_{out}(a_j). \tag{6.5}$$

In der Regel ist die Outputfunktion global definiert. Häufig, wie auch in diesem Modell, wird die **Identität** verwendet. Es gilt also

$$f_{\text{out}}(a_j) = a_j, \text{ also } o_j = a_j. \tag{6.6}$$

Die Informationsweitergabe erfolgt von Schicht zu Schicht, beginnend bei der Inputschicht, wobei jedes Neuron die erhaltenen Informationen wie in Abbildung 6.5 auf die gleiche Art und Weise verarbeitet.

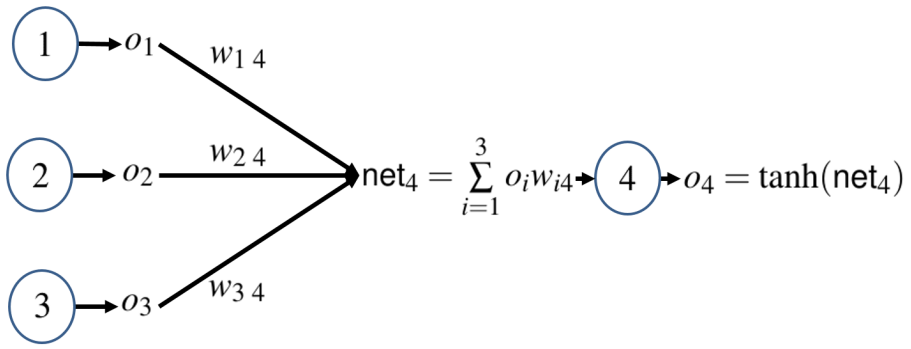


Abbildung 6.5: Schematische Darstellung der Informationsaufnahme und -weitergabe innerhalb eines einzelnen Neurons, welches kein Inputneuron ist

6.3.4 Variablen für Backpropagation

Tabelle 6.5 führt die Variablen für diesen Abschnitt ein.

Variable	Beschreibung
$\boldsymbol{x}$	Inputvektor eines Neuronalen Netzes nach den Variablen aus Abschnitt 3.2.2
$\boldsymbol{y}$	Outputvektor eines Neuronalen Netzes
I	Menge der Inputneurone
O	Menge der Outputneurone
W	Menge der Kantengewichte
$\Omega_j, j = (1, ..., \lvert O \rvert)$	Outputneuron
i	Ein Neuron
o	Output eines Neurons
P	Menge aller Trainingsbeispiele
p	Dem Netz präsentierter Input (Trainingsbeispiel), dazu existiert ti
ti	Der Teaching Input, welcher dem korrekten Output entspricht
M	Menge der Neuronen in einer Vorgängerschicht
L	Menge der Neuronen in einer Nachfolgerschicht
J	Menge der Neuronen in der aktuellen Schicht
j	Ein Neuron in der aktuellen Schicht
η	Lernrate, entspricht der Skalierung eines Anpassungsschrittes
t	Oberer Index, der eine Zeitabhängigkeit verdeutlicht
$\mathrm{Err}(W)$	Fehlerfunktion bezüglich der Menge der Gewichte
$g_{i,j}$	Gradient der Ableitung des Fehlers nach dem Gewicht der Kante (i, j)

Tabelle 6.5: Variablen dieses Abschnittes

Für spätere Zwecke sei bereits an dieser Stelle die Gleichung

$$g_{i,j} = \frac{\partial \mathrm{Err}(w)}{\partial w_{i,j}} = o_i \delta_j \tag{6.7}$$

gegeben.

6.3.5 Lernen: Backpropagation-Of-Error

In Abschnitt 6.3.3 wurde festgelegt, wie das Neuronale Netz einen Schluss aus einer Reihe von Inputinformationen zieht. Nun muss geklärt werden, wie das Netz derart angepasst wird, dass es auch die richtigen Schlüsse bildet. Diese Anpassung/Optimierung nennt sich **Lernen**. In diesem Fall **überwachtes Lernen**. Dabei erfolgt die Eingabe von Informationen in das ungelernte Netz, wobei der gewünschte Output stets bekannt ist. Es erfolgt eine Vorwärtspropagierung der Informationen durch alle Schichten. Anhand der Differenz von erwartetem und tatsächlichem Output (letzterer nennt sich **Teaching Input**) können Verbesserungen des Netzes berechnet werden. Es erfolgt die Anwen-

dung dieser Verbesserungen. Prinzipiell gibt es dafür folgende Möglichkeiten

- Entwicklung neuer Kanten;
- Entfernung vorhandener Kanten;
- Änderung der Kantengewichte;
- Änderung der Schwellwerte;
- Entfernung vorhandener Neuronen;
- Entwicklung neuer Neuronen;
- Änderung der Propagierungs- oder Outputfunktion.

Der verwendete Backpropagation-Algorithmus arbeitet mit der Änderung der Kantengewichte. Besitzt eine Kante den Wert 0, so gilt sie als nicht existent. Wird dieser Wert verändert, gilt sie als wieder aktiv. Damit deckt der Algorithmus theoretisch die ersten fünf Fälle ab. Der Schwellwert gehört auch dazu, da er aufgrund des Schwellwertneurons über seine Kantengewichte definiert wird. Die Bestimmung einer optimalen Anzahl von Neuronen und Schichten ist somit eine Frage des geschickten Testens. Die Änderungen der verwendeten Funktionen sind im Allgemeinen sehr schwierig zu realisieren und nicht intuitiv. Zudem gibt es dafür kaum eine biologische Motivation. Somit erscheint ein Algorithmus wie Backpropagation ideal.

Es existieren zwei beliebte Varianten des Lernens, genauer gesagt der Präsentation der Inputs. Diese sind das **Online** und das **Offline Lernen**. Beim Online Lernen ist das Netz nach jeder Inputauswertung zu ändern. Die Anpassung des Netzes geht dabei im Allgemeinen schneller vonstatten als das Online Lernen, bei dem erst alle Inputs ausgewertet werden und danach eine Anpassung vorgenommen wird. Der Zeitfaktor revidiert sich im nächsten Abschnitt. Der Nachteil des Online Lernens besteht darin, dass bereits erreichte Fortschritte durch einen einzigen Input zurückgenommen werden können. Die Offline Variante ist durch seine gleichzeitig verarbeitete Masse an Trainigsbeispielen wesentlich unempfindlicher gegenüber solchen Negationen. Für den Backpropagation-Algorithmus und vor allem für den Resilient-Backpropagation-Algorithmus ist die Offline Variante besser geeignet beziehungsweise die einzig anwendbare Möglichkeit.

Bemerkung 6.3.5
Im Folgenden werden alle nötigen Gleichungen hergeleitet. Für einen besseren Überblick gibt es am Ende des Abschnitts mit Abbildung 6.6 einen graphischen Überblick über die Zusammenhänge der hergeleiteten Formeln.

Der Backpropagation-Algorithmus ist ein gradientenbasiertes Optimierungsverfahren. Über die Optimierung der Fehler einer Netzeingabe werden die Korrekturwerte für die Kantengewichte zwischen der letzten und der vorletzten Schicht ermittelt. Anhand der veränderten Werte erfolgt die schrittweise Änderung der vorherigen Kantengewichte. Für die Herleitung des Backpropagation-Algorithmus ist der erste Teil bereits beschrieben. Die Propagierung der Informationen eines Trainingsbeispiels p durch das Netz wurde bereits in Abschnitt 6.3.3 beschrieben. Die anzupassende Fehlerfunktion für das Offline-Lernen lautet

$$\text{Err} : W \to \mathbb{R} \text{ mit } \text{Err}(W) = \frac{1}{2} \sum_{p \in P} \left(\sum_{\Omega \in O} \left(\text{ti}_{p,\Omega} - y_{p,\Omega} \right)^2 \right) \tag{6.8}$$

mit der spezifischen Fehlerfunktion für ein Trainingsbeispiel

$$\text{Err}_p(W) = \frac{1}{2} \sum_{\Omega \in O} \left(\text{ti}_{p,\Omega} - y_{p,\Omega} \right)^2. \tag{6.9}$$

Err nimmt W als Vektor entgegen, mit dem die Gewichtswerte auf den normalisierten Ausgabefehler abgebildet werden. Die Normalisierung erfolgt, um die Ausgabefehler in einem einzigen Wert $e \in \mathbb{R}$ abzubilden. Durch die Änderung der Gewichte wird nun versucht, das Minimum der Fehlerfunktion zu finden. Es gilt

$$\triangle W = -\eta \bigtriangledown \text{Err}(W).$$

Für ein einzelnes Gewicht von i nach j ergibt sich

$$\triangle w_{i,j} = -\eta \frac{\partial \text{Err}(W)}{\partial w_{i,j}} = \sum_{p \in P} -\eta \frac{\partial \text{Err}_p(W)}{\partial w_{i,j}}. \tag{6.10}$$

Mit den Gleichungen (6.2), (6.4) und (6.6) gilt

$$o_j = f_{\text{out}}(net_j) = \tanh\left(\sum_{i \in M} o_i w_{i,j} \right). \tag{6.11}$$

Durch die Funktionsschachtelung ist die Kettenregel auf die *spezifische Fehlerfunktion* in Gleichung (6.10) anwendbar. Die letztendlich ermittelte Gleichung ist damit für jedes Trainingsbeispiel gleich. Der Index p zur Kennzeichnung eines speziellen Trainingsbeispiels entfällt ab sofort zu Gunsten der besseren Lesbarkeit. Damit es nicht zu Verwechslungen kommt, wird $\text{Err}_p(W)$ mit Err abgekürzt und ist nicht mit $\text{Err}(W)$ zu verwechseln. Eine schematische Darstellung der vielfachen Anwendungen der Kettenregel ist am Ende der Herleitung in Abbildung 6.6 gegeben. Die Anwendung der Kettenregel

bezüglich Gleichung (6.11) liefert

$$\frac{\partial \text{Err}}{\partial w_{i,j}} = \frac{\partial \text{Err}}{\partial \text{net}_j} \cdot \frac{\partial \text{net}_j}{\partial w_{i,j}}. \tag{6.12}$$

Für den zweiten Faktor gilt

$$\frac{\partial \text{net}_j}{\partial w_{i,j}} = \frac{\partial \sum\limits_{i \in M} o_i w_{i,j}}{\partial w_{i,j}} = \boxed{o_i}. \tag{6.13}$$

Für den ersten Term sei die Bezeichnung $-\delta_j$ eingeführt. Die Kettenregel wird mit Gleichung (6.11) auf δ_j angewandt

$$\delta_j = -\frac{\partial \text{Err}}{\partial \text{net}_j} = -\frac{\partial \text{Err}}{\partial o_j} \cdot \frac{\partial o_j}{\partial \text{net}_j}. \tag{6.14}$$

Der zweite Faktor liefert aufgrund der Gleichungen (6.11) und (6.11) die Ableitung der Outputfunktion

$$\frac{\partial o_j}{\partial \text{net}_j} = \frac{\partial \tanh(\text{net}_j)}{\partial \text{net}_j} = \boxed{1 - \tanh^2(\text{net}_j)}. \tag{6.15}$$

An dieser Stelle kommt der Begriff des *Back*propagation, des Rückwärtsrechnens zum Tragen, denn die erste Derivate der Fehlerfunktion nach dem Output einer Neuronenschicht J steht in Abhängigkeit zum Vektor aller Netzeingaben der Vorgängerschicht. Entsprechend kann die nachfolgenden Gleichung (6.16) abgewandelt und die Kettenregel angewandt werden

$$-\frac{\partial \text{Err}}{\partial o_j} = \frac{\partial \text{Err}\,(\text{net}l_1), \text{net}l_l)}{\partial o_j} = \sum_{l \in L} \left(-\frac{\partial \text{Err}}{\partial \text{net}_l} \cdot \frac{\partial \text{net}_l}{\partial o_j} \right). \tag{6.16}$$

Eine exakte Rechenvorschrift entsteht für den zweiten Faktor, wenn net_l durch seine Summendarstellung aus Gleichung (6.2) ersetzt wird

$$\frac{\partial \text{net}_l}{\partial o_j} = \frac{\partial \sum\limits_{j \in J} o_j w_{j,l}}{\partial o_j} = \boxed{w_{j,l}}. \tag{6.17}$$

Die Form des ersten Faktors ist bereits aus Gleichung (6.14) bekannt, nämlich

$$-\frac{\partial \text{Err}}{\partial \text{net}_l} = \boxed{\delta_l}. \tag{6.18}$$

Nach Zusammensetzung aller umrahmten Terme ergibt sich die **Verallgemeinerte Delta-Regel**, genannt **Backpropagation of Error** für das Offline-Lernen

$$\triangle w_{i,j} = \eta \sum_{p \in P} o_{p,i} \delta_{p,j} \text{ mit}$$
$$\delta_{p,j} = \begin{cases} \left(1 - \tanh^2(\text{net}_{p,j})\right) \left(\text{ti}_{p,j} - y_{p,j}\right) \forall j \in O \\ \left(1 - \tanh^2(\text{net}_{p,j})\right) \sum\limits_{l \in L} \left(\delta_{p,l} w_{p,j,l}\right). \end{cases} \tag{6.19}$$

Der gesamte Algorithmus mit den Verbesserungen des Resilient-Backprop-Algorithmus wird in Abschnitt 6.3.8 angegeben.

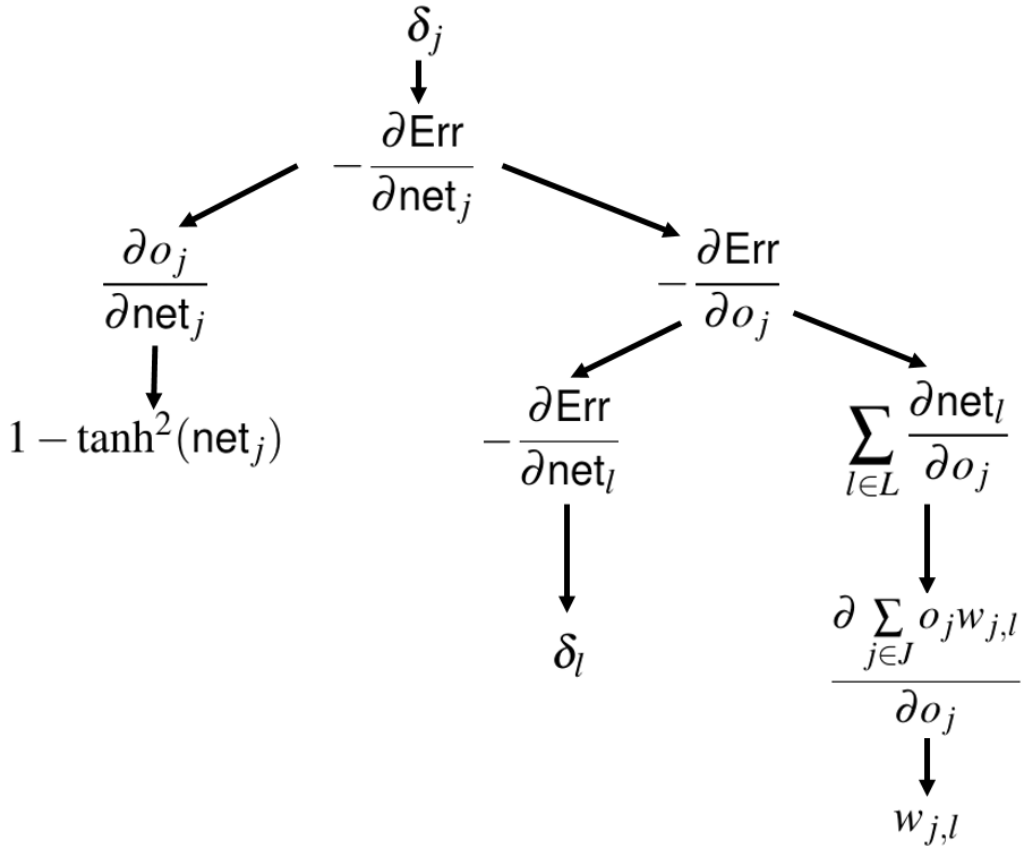


Abbildung 6.6: Schematische Darstellung der mehrfachen Anwendung der Kettenregel bei den Derivativen von δ, beginnend bei δ_j.

6.3.6 Ein Wort zu den Schichten

Im vorherigen Abschnitt wurde bereits erwähnt, dass diese Implementation des Backpropagation-Algorithmus mit drei inneren Neuronenschichten zu je p Neuronen arbeitet. Im Allgemeinen gilt: Je mehr Schichten und je mehr Neuronen pro Schicht existieren, desto angepasster die Lösung. Es nimmt also die Varianz zu und der Bias ab. Allerdings ist dabei stets zu beachten, dass „mehr“ auch mehr Rechenzeit und Input bei weniger zusätzlicher Leistung bedeutet. Zugleich droht eine Überanpassung. Es bedarf einer großen Menge an Trainingsdaten, wenn die Anzahl der inneren Neuronen hoch ist, wobei mit hoch schon eine mittlere zweistellige Anzahl gemeint sein kann. Des Weiteren gilt, dass mehr als eine notwendige Anzahl an Schichten meist zu keiner sonderlichen Verbesserung der Güte des Ergebnisses führen. Zusätzlich steigt die Rechenzeit exponentiell mit der Anzahl der Schichten. In der Praxis hat sich die Verwendung von drei inneren Neuronenschichten durchgesetzt, weshalb diese Menge für das Modell übernommen wurde. Die Anzahl der inneren Neuronen ist den p Inputparametern nachempfunden und hat sich als praktikabel erwiesen.

Die Anzahl der Neuronenschichten entstammt aus einem Ursprungsproblem der Neuronalen Netze: Der Separierung von Datenpunkten in einem Koordinatensystem. Anfänglich existierten zwei Arten von Datenpunkten, die sich zum Beispiel in einer binärcodierten Eigenschaft, wie zum Beispiel einer Farbe, unterschieden. Diese Punkte liegen in einer Ebene und sollen mittels eines Algorithmus separiert werden. Das Neuronale Netz liest die Eigenschaften eines Punktes ein und gibt seine Zugehörigkeit zu den Farblagern aus. Bildlich gesprochen entstehen trennende Hyperebenen im Koordinatensystem.
Der Physiker John von Neumann [ELE] behauptete, mit vier Parametern könne er einen Elefanten darstellen. Mit einem fünften wäre er in der Lage, das Wackeln seines Schwanzes zu simulieren.
Frei nach diesem Motto kann mit Hilfe *einer* inneren Neuronenschicht eine separierende Gerade in einem zweidimensionalen Koordinatensystem erzeugt werden. Zwei Schichten führen zur Bildung konvexer Polygone und bei drei Schichten ist die Modellierung beliebiger und beliebig vieler Mengen möglich. Siehe dazu untenstehende Abbildung 6.7. Die Anwendung ist in den mehrdimensionalen Raum übertragbar.

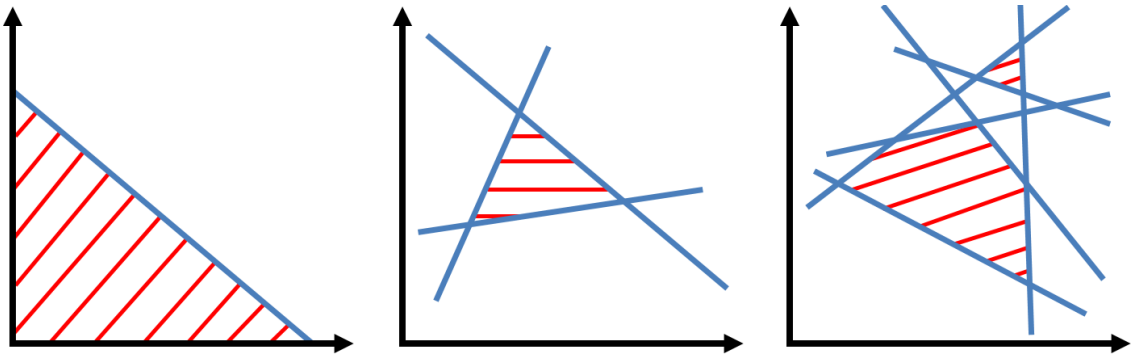


Abbildung 6.7: Schematische Darstellungsmöglichkeiten separierender Geraden/Mengen in einem zweidimensionalen Koordinatensystem mit: Links) einer Schicht, Mitte) zwei Schichten, Rechts) drei Schichten

.

6.3.7 Nachteile des Backpropagation-Algorithmus

Bei der Anwendung von Gradientenverfahren gibt es im Allgemeinen vier Nachteile, die je nach Wahl der Daten, der Einstellungen, des Verfahrens und der Startlösung eintreten können. Diese vier Fälle sind in Abbildung 6.8 aufgezeigt.

1. Es ist im Allgemeinen davon auszugehen, dass komplexe Zielfunktionen mehr als nur ein Minimum enthalten. Da das Gradientenverfahren von einer Zwischenlösung lediglich in die *lokal* beste Richtung voranschreitet, kann das Verfahren in einem schlechten lokalen Minima enden. In der Regel ist nicht bekannt, ob eine bessere Lösung existiert.

2. Beim Durchlaufen eines flachen Plateaus wird der Gradient zwangsweise sehr klein. Zum Verlassen des Abschnittes sind oft sehr viele Rechenschritte notwendig, wenn es keine automatische Anpassung der Schrittweite gibt.

3. Bei einer zu großen Schrittweite können gute Minima wieder verlassen werden. In diesem Fall sorgt eine automatische Schrittweitensteuerung dafür, dass der nächste Schritt bei großem Gefälle geringer ausfällt. Eine Garantie besteht jedoch nicht.

4. Wechselt der Gradient in einer steilen Schlucht sprunghaft von einem stark positiven zu einem stark negativen Wert, so kann der seltene Fall der Oszillation eintreten. Dabei springt der Algorithmus zu einem bereits bearbeiteten Punkt zurück, sodass sich das Verfahren in einer Endlosschleife verfängt.

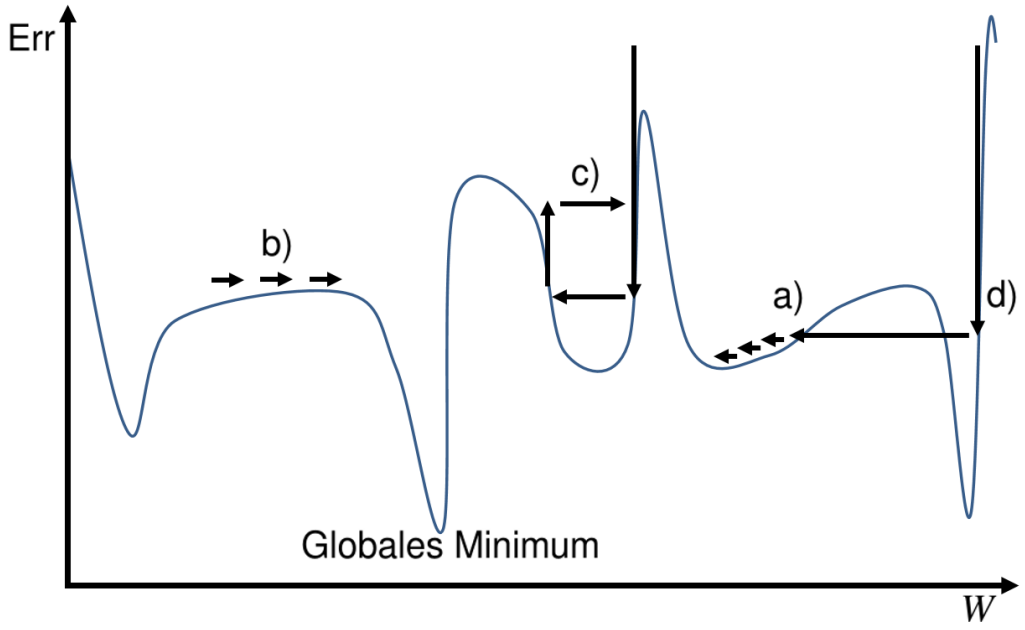


Abbildung 6.8: Beispiel für die Nachteile eines Gradientenverfahrens: a) Finden schlechter Minima, b) Quasi-Stillstand bei kleinen Gradienten, c) Oszillation in Schluchten, d) Verlassen guter Minima Die Abbildung ist [KRI05] nachempfunden.

Die Erweiterung des Backpropagation-Algorithmus, der Resilient-Backprop-Algorithmus enthält veränderte Formeln und Funktionen, sodass die Gefahr des Eintretens einiger der eben beschriebenen Nachteile um einen entscheidenden Faktor reduziert werden kann. Namentlich sind dies Nummer 1, 2 und 3.

6.3.8 Lernen: Resilient-Backpropagation und andere Anpassungen

Die im letzten Abschnitt aufgeführten Nachteile sollen durch den hier vorgestellten Algorithmus entschärft werden. Die von den deutschen Forschern Martin Riedmiller und Heinrich Braun 1993 erdachte Weiterentwicklung des Backpropagation-Algorithmus, der Resilient Backpropagation-Algorithmus (kurz: „Rprop“) benutzt die Lernrate η zur Anpassung der Schrittweite. Die Größe des Gradienten spielt keine Rolle mehr. Ein weiteres, bisher noch nicht genanntes Problem besteht in der verlangsamten Laufzeit des Algorithmus bei Berechnungen in einer Neuronenschicht, die weit von der Inputschicht entfernt ist. Eine Lösung dieses Problems lässt sich in der Änderung der bis dato vom Benutzer fest vorgegebenen Lernrate finden.

Anpassung der Lernrate η

Bisher kam beim Backpropagation-Algorithmus eine mehr oder weniger fest vorgegebene Lernrate η zum Einsatz. Sie konnte auch je nach Schicht einzeln definiert sein. Der

Ansatz von Rprop geht davon aus, dass keine global gültige Lernrate eingesetzt wird, sondern jedes Gewicht $w_{i,j}$ eine spezifische Lernrate $\eta_{i,j}$ besitzt, welche so in den Algorithmus eingebettet ist, dass sie sich während der Berechnung automatisch anpasst. Zur Verdeutlichung der Zeitabhängigkeit kann mit $\eta_{i,j}^{(t)}$ gearbeitet werden. Damit wird einerseits das Problem des schichtweise verlangsamten Lernens behoben, als auch ein gezielteres Lernen ermöglicht.
Zur Anpassung der Lernraten $n_{i,j}$ sind die Gradienten aus dem aktuellen Zeitschritt $g_{i,j}^{(t)}$ und dem vergangenen Zeitschritt $g_{i,j}^{(t-1)}$ zu betrachten. Die Größe der Gradienten ist dabei nicht von Interesse. Es interessiert einzig und allein das Vorzeichen.
Bei wechselndem Vorzeichen ist ein lokales Minimum der Fehlerfunktion übersprungen worden. Dies bedeutet, dass der in $(t-1)$ ausgeführte Anpassungsschritt zu groß war. Er ist also mit einer geringeren Lernrate zu wiederholen. Dafür muss $\eta_{i,j}^{(t-1)}$ mit einer Konstante $\eta_{\downarrow} \in [0,1]$ multipliziert werden. Die Nicht-Durchführung des Gewichtsupdates ist implementationstechnisch durch die Festlegung des Gewichtes $w_{i,j}^{(t)}$ auf 0 zu realisieren.
Bei gleichbleibenden Vorzeichen kann eine leichte Vergrößerung von $\eta_{i,j}$ vorgenommen werden. Dies hat zur Folge, dass die Überwindung flacher Bereiche schneller vonstatten geht, als beim Backpropagation-Algorithmus. Dazu muss $\eta_{i,j}^{(t-1)}$ mit einer Konstante $\eta_{\uparrow} \in (1,\infty)$ multipliziert werden.

Definition 6.3.9 (Anpassung der Lernrate in Rprop)

$$\eta_{i,j}^{(t)} = \begin{cases} \eta_{\uparrow}\eta_{i,j}^{(t-1)}, \textit{ falls } g_{i,j}^{(t-1)}g_{i,j}^{(t)} > 0 \\ \eta_{\downarrow}\eta_{i,j}^{(t-1)}, \textit{ falls } g_{i,j}^{(t-1)}g_{i,j}^{(t)} < 0 \\ \eta_{i,j}^{(t-1)} \textit{ sonst.} \end{cases} \tag{6.20}$$

Dies hat zur Folge, dass Rprop ein Offline Verfahren ist. Ohne die Kontinuität der Gradienten folgt eine Verlangsamung des Verfahrens auf niedrigstes Niveau. Die automatische Anpassung ermöglicht eine Verfahrensbeschleunigung auf eine Ebene, die mit der Online-Variante des Backpropagation-Algorithmus mindestens gleichzusetzen ist.

Laut Kriesel [KRI05] hat sich ein Initialisierungswert von $\eta_{i,j}^{(0)} = 0 \; \forall i,j$ bewährt. Diese oder andere Werte seien aufgrund der schnellen Überschreibung unkritisch zu sehen, solange sie positiv und nicht übertrieben groß angesetzt sind.
Für η_{max} ist der als unkritisch anzusehende Wert 50 empfohlen. Niedrigere Maximalwerte führen zu kleineren Updateschritten. Eine geringe Schrittweite sollte zu jeder Zeit ermöglicht werden. Damit ist ein Wert von $\eta_{\text{min}} = 10^{-6}$ definitiv vertretbar.
Wie bereits gesagt, bedeutet die Anwendung von $\eta_{\downarrow}$ das Überspringen eines Minimums mit unbekannter Position. Für eine systematische Suche danach bietet sich logischerweise $\eta_{\downarrow} = 0.5$ an. Für $\eta_{\uparrow}$ hat sich ein Wert von 1.2 als effizient erwiesen, wobei leichte Änderungen keinen signifikanten Einfluss auf die Konvergenzgeschwindigkeit besitzen. Damit kann $\eta_{\downarrow} = 1.2$ als Konstante festgelegt werden.

Anpassung der Gewichtsänderung

Die Anpassung der Gewichte erfolgte im Backpropagation-Algorithmus proportional zum Gradienten der Fehlerfunktion. Damit wird das vollständige Relief der Fehlerfunktion übernommen. Die Effizienz dieser Maßnahme darf angezweifelt werden. Der Betrag der Gewichtsänderung $\triangle w_{i,j}$ beim Rprop hingegen entspricht direkt der im vorherigen Abschnitt 6.3.8 angepassten Lernrate $\eta_{i,j}$. Damit entfällt die Proportionalität der Gewichtsänderung zum Gradienten. Lediglich sein Vorzeichen geht in die Anpassung ein. Das Aussehen des Ergebnisprozesses ist dadurch weniger zerklüftet. Dadurch obliegt dem Gradienten nur noch die Richtung, aber nicht die Stärke des Abstieges.
Bei positivem Vorzeichen von $g_{i,j}$ ist das Gewicht $w_{i,j}$ zu verringern. Dies wird durch die Subtraktion von $\eta_{i,j}$ von $w_{i,j}$ erreicht.
Bei negativem Vorzeichen erfolgt die Addition von $\eta_{i,j}$ zu $w_{i,j}$.
Ein Gradient mit dem Wert 0 bedarf keiner Anpassung.

Definition 6.3.10 (Gewichtsänderung in Rprop)

$$\triangle w_{i,j}^{(t)} = \begin{cases} -\eta_{i,j}^{(t)}, \textit{ falls } g_{i,j}^{(t)} > 0 \\ +\eta_{i,j}^{(t)}, \textit{ falls } g_{i,j}^{(t)} < 0 \\ 0 \textit{ sonst} \end{cases} \tag{6.21}$$

Eine Erweiterung der Gewichtsänderung bietet der sogenannte **Momentum-Term**

$$\triangle w_{i,j}^{(t)} = \eta o_i \delta_j + \alpha \cdot \triangle w_{i,j}^{(t-1)} \text{ mit } \alpha \in [0.6, 0.9]\,. \tag{6.22}$$

Gleich einem Skifahrer sorgt der Term dafür, dass ein Teil der Bewegung aus dem vorherigen Schritt in den nächsten übernommen wird. Damit kann das Problem des Verweilens in schlechten lokalen Minima gemildert werden, da eine Überspringung derselben möglich ist. Dafür ist die Wahrscheinlichkeit des Verlassens guter Minima erhöht. Dieser Term wurde in Rprop-angepasster Form implementiert. Andere Anpassungsmöglichkeiten sind beispielsweise **Flat-Spot-Elimination**, **Second-Order-Backpropagation**, **Weight-Decay**, **Pruning** und **Optimal Brain Damage**.

Der Algorithmus

Im Folgenden wird der Resilient-Backpropagation-Algorithmus angegeben.

Algorithm 2 Resilient-Backpropagation-Algorithmus

Initialisiere das Netz mit $w_{i,j} = 1$, $\eta_{i,j} = 0.1$, $\eta_{\text{min}} = 10^{-6}$, $\eta_{\text{max}} = 50$, $\eta_{\downarrow} = 0.5$, $\eta_{\uparrow} = 1.2$, $g_{i,j}^{(t)} = 1$, $t = 0$
Wähle Err(W) zu groß
while Err(W) zu groß **do**
  $t = t + 1$

  Es folgt die Berechnung der Vorwärtspropagierung:
  for $p \in P$ **do**
    for $i \in I$ **do**
      o_i ist der i-te Inputparameter
    end for
    for J ist innere Schicht $1, 2, 3$ und äußere Schicht O **do**
      M ist Vorgängerschicht
      for $j \in J$ **do**
        $\text{net}_j = \sum_{i \in M} o_i w_{i,j}$
        $o_j = \tanh(\text{net}_j)$
      end for
    end for
  end for

  Berechne Err$(W) = \frac{1}{2} \sum_{p \in P} \sum_{j \in O} (o_j - t_j)^2$

  Es folgt die Berechnung der Rückwärtspropagierung:
  for J in äußere Schicht O und innere Schicht 3,2,1 **do**
    i ist ein Neuron der Vorgängerschicht M, l ein Neuron der Nachfolgerschicht L
    for $j \in J$ **do**
      Bilde $\eta_{i,j}^{(t)}$ nach Regel (6.20)
      Bilde $\triangle w_{i,j}^{(t)}$ nach Regel (6.21)
      Mit $g_{i,j} = o_i \delta_j$, δ_j nach Regel (6.7)
      Erweitere $\triangle w_{i,j}^{(t)}$ mit dem Momentum-Term $M \cdot \triangle w_{i,j}^{(t-1)}$
      Mit $M = 0.1$
    end for
  end for
end while

Bemerkung 6.3.6
Um anschließend einen Entscheidungswert für eine konkrete Route zu berechnen, ist lediglich die Vorwärtspropagierung bei angepasstem Netz für diesen einen Routeninput durchzuführen. Das Ergebnis der Outputschicht liefert den Entscheidungswert.

6.3.9 Vor- und Nachteile

Der Resilient-Backpropagation-Algorithmus als Gradientenabstiegsverfahren bietet den Vorteil, dass er lediglich zwei Eingabemonate benötigt, um Vergleichsdaten zu erstellen. Somit kann er zeitlich nah am vorauszusagenden Monat gehalten werden.
Des Weiteren ist seine Laufzeit im Allgemeinen besser als der gewöhnliche Backpropagation-Algorithmus.
Im Allgemeinen bestehen die in Abschnitt 6.3.7 aufgeführten Nachteile, welche größtenteils durch den Resilient-Backpropagation-Algorithmus gemildert sind.

Die Aufgabe des Neuronalen Netzes besteht in der Entscheidung der Existenz oder Nichtexistenz einer Route bei Eingabe ihrer Parameter. Vormodell 1 muss erneut durchgeführt werden, um topografisch unabhängige Daten zu ermitteln. Das Neuronale Netz ist aber bezüglich der Anzahl der Ausgabeneuronen erweiterbar, sodass zusätzlich zum Entscheidungswert der Existenz die durch Vormodell 1 bestimmten Daten berechnet werden können.

Am Minimum der lokalen Approximation herrscht eine Unstetigkeitsstelle. Daher besteht die Gefahr des Überspringens des Minimums, welche durch die Einführung des Momentum-Terms noch erhöht wird. Dennoch ist dieser Term enthalten, um mögliche schlechte Minima zu überspringen.
Zur Erhöhung der Stabilität und Geschwindigkeit haben Christian Igel und Michael Hüsken im Jahr 2003 einige leichte Änderungen am Algorithmus vorgenommen [IGE]. Ebenso nahmen A.D. Anstasiadis, G.D. Magoulas und M.N. Vrahatis einige Abwandlungen vor, sodass der Algorithmus global konvergiert [ANA]. Diese Änderungen konnten aufgrund zeitlicher Umstände nicht mehr implementiert werden.

7 Auswertung

In diesem Kapitel sollen die verschiedenen Modelle miteinander verglichen und eine Wertung erstellt werden. Grundlage des Vergleichs bilden Fehlermaße und Kennzahlen, welche in den Tabellen 7.3 und 7.4 mittels Bezeichnung und Berechnungsformel aufgeführt sind. Die benötigten Variablen stehen in 7.1 und Tabelle 7.2 liefert die Berechnungsformeln der wichtigsten Werte. Argumentationsansätze und Werte wurden [HÖF04], [BRO81], [IPDS], [KOELN], [LST], [INWT], [WWW] und [CKS] entnommen.

7.1 Fehlermaße

Ein Vergleich der Modelle mittels üblicher Kennwerte wie Akaikes Informationskriterium, dem Theilschen Ungleichheitskoeffizienten oder dem Gini-Index scheidet aufgrund der unterschiedlichen Modellansätze aus. Das naive Modell besitzt gar keinen Ansatz, das lineare Modell und der lineare Bootstrap besitzen einen linearen, die logistische Regression und der Conditional Logit einen nichtlinearen und das Neuronale Netz mit seinem Perceptron Modell bildet sogar eine eigene Klasse. Die für die verschiedenen Modelle eingesetzten Gütekriterien sind somit größtenteils nicht anwendbar. Als Basis für Vergleiche dient damit nur der von den Modellen ermittelte Output in Form der Differenz der vorhergesagten und tatsächlichen Werte. Diese Gegenüberstellung der geschätzten und tatsächlichen Werte eines Validierungsdatensatzes nennt sich **Ex-Post-Evaluierung**.

Die verwendeten Maße lassen sich grob in normierte, quadratische, relative und einfache Fehler einteilen.
Die einfachen Maße wie ME bestimmen einfach nur die Differenz aus prognostizierten und wahren Outputs. Sie werden für Vorhersagen mit verschiedenen Werteniveaus verwende, sind aber auch kritisch zu betrachten. Eine gute Prognosegenauigkeit kann aus sich ausgleichenden Fehlerwerten hervorgehen. Normierte Maße wie MAE vermeiden dieses Problem und sind zudem leichter interpretierbar.
Die quadratischen Maße wie MSE betrachten dieselben Differenzen wie die einfachen Maße, allerdings gehen sie quadratisch in die Wertung ein. Neben der Normierung besteht der Hauptgedanke darin, zu erkennen, ob sehr starke Ausreißer oder tendenziell hohe Abweichungen in den Prognosedaten vorhanden sind. Durch Entfernung der Ausreißer und einer erneuten Berechnung lässt sich erkennen, ob lediglich einige sehr große Ausreißer in den Werten waren (dann ist der zweite Berechnungswert deutlich kleiner als der erste) oder ob die geschätzten Werte tendenziell hohe Abweichungen produzieren. Dann ist das Schätzmodell als ungeeignet einzustufen. Sein eigentlicher Vorteil liegt aber in seiner Differenzierbarkeit nach der Substitution der Outputvariablen mit der Berechnungsvorschrift eines Schätzmodells. Danach bildet der MSE die zu op-

timierende Fehlerfunktion des „Kleinste-Quadrate-Anpassungsmodells“.

Zu jedem der eben erwähnten Fehler lässt sich eine relative Variante berechnen, welche die Abweichung im Verhältnis zum wahren Wert angibt. So zum Beispiel der MPE und der MAPE. Damit lässt sich grob gesagt die prozentuale Abweichung angeben und liefert anschauliche Werte, ohne dass Maßeinheiten zu berücksichtigen sind.
Bei derartigen Werten eignet sich die Ermittlung des Medians. Er bildet eine Hilfe, um zu erkennen, ob die Verteilung der prozentualen Fehler tendenziell linksschief, rechtsschief oder symmetrisch ist. Eine zur 0 neigende Linksschiefe ist prinzipiell gut, da sie das Vorhandensein vieler kleiner prozentualer Abweichungen andeutet.
Weiterhin seien die Wurzelvarianten der quadratischen Fehler erwähnt, so zum Beispiel der RMSE. Da er lediglich die Wurzel des MSE ist, besitzt er dieselben Eigenschaften und denselben Wert nur etwas kleiner. Sein Vorteil ist, dass er dieselbe Maßeinheit besitzt wie die Outputwerte. Hier wird der MSE anstelle des RMSE angegeben, da ersterer Wert von anderen Fehlermaßen verwendet wird und er somit interessanter ist, als der RMSE.

Es verbleiben die Gütemaße aus Tabelle 7.4, welche zu den normierten Fehlermaßen zu zählen sind. Sie reagieren allerdings nicht gleich sensitiv auf die Wertedifferenzen, sondern besitzen ihre eigene Aussagekraft.
Die Wirkung von großen oder kleinen Werten des Bias- und Varianz-Anteils des MSE wurden für den Bias und die Varianz allein bereits in Abschnitt 4.4 beschrieben. Sie geben den im MSE enthaltenen Modellfehler (Bias) und seine Empfindlichkeit gegenüber Änderungen in den Eingabedaten (Varianz) wieder.
Ein wichtiges Maß bildet der Bravais-Pearsonsche Korrelationskoeffizient BPC $\in [-1, 1]$. Für zwei Variablen zeigt er an, ob diese in einem linearen Zusammenhang stehen oder nicht. Er ist nicht ausschließlich für lineare Regressionen anwendbar und zeigt ebenso nichtlineare Zusammenhänge an.

- BPC $= 0$: Die prognostizierten und beobachteten Outputwerte stehen in nichtlinearem Zusammenhang.
- BPC $= 1$: Die prognostizierten und beobachteten Outputwerte stehen in linearem Zusammenhang.
- BPC $= -1$: Die prognostizierten und beobachteten Outputwerte stehen in negativ linearem Zusammenhang.

Aus der linearen Unkorreliertheit (BPC $= 0$) folgt nicht die Unabhängigkeit.
Des Weiteren existiert der Determinationskoeffizient $R^2 \in [0, 1]$, auch Bestimmtheitsmaß genannt [WEB63]. Er entspricht dem Anteil der Varianz des geschätzten Outputs an der Varianz des wahren Outputs der Beobachtungen. Mit ihm soll ermittelt werden, ob die gewählten Inputvariablen eines linearen Regressionsmodelles geeignet sind den Output vorherzusagen. Für die logistische Regression und das Neuronale Netz wird er vermutlich wenig aussagekräftig sein, beziehungsweise den nichtlinearen Zusammen-

hang bestätigen.

- $R^2 = 0$: Der Input lässt keine Prognose des Outputs zu.
- $R^2 = 1$: Das Modell ist perfekt angepasst.

1-R^2 beziffert dementsprechend den nicht gemeinsamen Anteil. Als Quadrat des Korrelationskoeffizienten liefert R^2 annähernd prozentual verwertbare Aussagen zum Zusammenhang zwischen Wirklichkeit und Prognose.

Ein weiteres nützliches Analysewerkzeug bieten die sogenannten Boxplots. Dabei handelt es sich um die graphische Darstellung der Verteilung der Elemente einer Menge. Im vorliegenden Fall soll visuell veranschaulicht werden, wie sich die Differenzen aus geschätztem und beobachtetem Output verteilen. Gedanklich ist ein Boxplot als Anzahl von Markierungen auf einem Zahlenstrahl vorstellbar. Der Zahlenstrahl beinhaltet die reellen Zahlen. Eine Eintragung entspricht dem Wert einer Differenz aus geschätztem und beobachtetem Output. Fünf Markierungen werden auf ihm eingetragen:

- Das 0.25 und das 0.75 Quantil. Innerhalb dieses Intervalls befinden sich 50% der Werte.
- Der Median der Werte. Er gibt an in welchem Bereich des Quantilintervalls sich jeweils 25% der Werte befinden und zeigt eine linksschiefe, rechtsschiefe oder symmetrische Tendenz an.
- Die Whisker. Sie stellen eine Erweiterung des Quantilintervalls dar und besitzen meist den 1.5- oder 3-fachen Quantilwert. Datenpunkte außerhalb des Whiskerintervalls werden als Ausreißer bezeichnet. Punkte innerhalb des 1.5- und 3-fachen Quantilwertes heißen „Milde Ausreißer“, Punkte außerhalb des 3-fachen Quantilwertes heißen „Extreme Ausreißer“.

Variable	Erklärung
n	Anzahl aller geschätzter Routen
$i = 1, ..., n$	Eine spezielle Route
$\widehat{y_i}$	Geschätzte Passagierzahl der Route i
y_i	Wahre Passagierzahl der Route i
$\boldsymbol{y}$	$1 \times n$-Vektor der wahren Routenpassagierzahlen

Tabelle 7.1: Tabelle der für die Fehlermaße verwendeten Variablen

Kürzel	Formel	Name
$\mu_{\widehat{y}}$	$\frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i$	Erwartungswert des geschätzten Outputs
μ_y	$\frac{1}{n}\sum_{i=1}^{n}y_i$	Erwartungswert des wirklichen Outputs
$\sigma^2_{\widehat{y}}$	$\frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i^2-\left(\mu_{\widehat{y}}\right)^2$	Varianz des geschätzten Outputs
σ^2_y	$\frac{1}{n}\sum_{i=1}^{n}y_i^2-(\mu_y)^2$	Varianz des wahren Outputs
$\sigma_{\widehat{y}}$	$\sqrt{\sigma^2_{\widehat{y}}}$	Standardabweichung des geschätzten Outputs
σ_y	$\sqrt{\sigma^2_y}$	Standardabweichung des wahren Outputs

Tabelle 7.2: Tabelle mit wichtigen Werten

Kürzel	Formel	Name
ME	$\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)$	Mittlerer Fehler
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert\widehat{y}_i - y_i\rvert$	Mittlerer absoluter Fehler
MPE	$100\% * \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{y}_i - y_i}{y_i}$	Mittlerer prozentualer Fehler
MAPE	$100\% * \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert\widehat{y}_i - y_i\rvert}{y_i}$	Mittlerer absoluter prozentualer Fehler
MdAPE	$100\% * \text{Median}\left(\frac{\lvert\widehat{y}_i - y_i\rvert}{y_i}\right)$	Median des absoluten prozentualen Fehlers
MSE	$\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2$	Mittlerer quadratischer Fehler
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}$	Wurzel des mittleren quadratischen Fehlers
RMSPE	$100\% * \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(\widehat{y}_i - y_i)^2}{y_i^2}}$	Wurzel des mittleren quadratischen prozentualen Fehlers

Tabelle 7.3: Tabelle der untersuchten Fehlermaße

Kürzel	Formel	Name
BIASMSE	$\frac{\left(\mu_{\hat{y}} - \mu_y\right)^2}{\text{MSE}}$	Bias-Anteil des MSE
VARMSE	$\frac{\left(\sigma_{\hat{y}} - \sigma_y\right)^2}{\text{MSE}}$	Varianz-Anteil des MSE
R^2	$\frac{\left(\sum_{i=1}^{n}\left(\left(\hat{y}_i - \mu_{\hat{y}}\right)\left(y_i - \mu_y\right)\right)\right)^2}{\left(\sum_{i=1}^{n}\left(\hat{y}_i - \mu_{\hat{y}}\right)^2\right)\left(\sum_{i=1}^{n}\left(y_i - \mu_y\right)^2\right)}$	Determinationskoeffizient
KOVMSE	$\frac{2(1-\text{R})\sigma_{\hat{y}}\sigma_y}{\text{MSE}}$	Kovarianz-Anteil des MSE
BPC	$\frac{\sum_{i=1}^{n}\left(\left(\hat{y}_i - \mu_{\hat{y}}\right)\left(y_i - \mu_y\right)\right)}{\sqrt{\left(\sum_{i=1}^{n}\left(\hat{y}_i - \mu_{\hat{y}}\right)^2\right)\left(\sum_{i=1}^{n}\left(y_i - \mu_y\right)^2\right)}}$	Bravais-Pearsonsche Korrelationskoeffizient

Tabelle 7.4: Tabelle der untersuchten Fehlermaße

7.2 Referenzzeitpunkt der Auswertung

Die Auswertung der Hauptmodelle und der Vormodelle erfolgt für die ermittelten Optimaleinstellungen zur Vorhersage der Routenpassagiere des Monats April 2014. Die Vorgabe des zeitlichen Abstandes des frühesten verwendeten Monats liegt bei mindestens einem Jahr. Dadurch sollen realistische Rahmenbedingungen simuliert werden. Der gewählte Referenzzeitpunkt lautet demzufolge April 2013. Testeinstellungen waren zumeist der zeitliche Abstand der verwendeten Zeitpunkte, es wurden also nur jährlich oder monatlich aufeinanderfolgende Zeitpunkte verwendet. Weiterhin wurden verschiedene Anzahlen der Zeitpunkte getestet. Bei Modellen wie dem linearen Bootstrap war zudem die Anzahl der Datensätze und Wiederholungen interessant.
Der anschließende Einstellungsvergleich erfolgte in Hinblick auf den RMSPE, den MPE und die benötigte Zeit. Eine Untersuchung hinsichtlich des in Abschnitt 4.4 erwähnten Bias-Varianz-Dilemmas erfolgte nicht, da sich die Werte in den Modellen meist um maximal 5 Prozentpunkte unterschieden. Hauptentscheidungskriterium war der gewählte

Referenzzeitpunkt bei dem allgemein gilt: Je näher er am zu schätzenden Zeitpunkt liegt, desto besser sind die Werte der Ergebnisse. Der Unterschied des Referenzzeitpunktes März 2014 statt April 2013 für den zu schätzenden Zeitpunkt April 2014 machte sich in einem grundlegenden Unterschied von mindestens 10 bis zu 15 Prozentpunkten bemerkbar.

Für das naive Modell besteht die Optimaleinstellung darin, lediglich den Referenzzeitpunkt zu verwenden und auch keine weiteren Routen aus älteren Zeitpunkten zu nutzen. Die Ergebnisse des linearen Modells liefern das übereinstimmende Ergebnis, das Optimaleinstellung sowohl bei den jährlichen, als auch bei den monatlichen Daten herrscht, insofern die verwendeten Routen aus den zurückliegenden zwei Zeitpunkten stammen (zuzüglich des Referenzzeitpunktes). Die *Werte* für die Routen dieser zwei Jahre (April 2012 und 2011) können aber auch aus weiter zurückliegenden Zeitpunkten stammen. Mit vier derartigen Zeitpunkten liefert die monatliche Einstellung hinsichtlich des MPE den besten Wert.
Die Einstellung des linearen Modells ist für den linearen Bootstrap übernommen worden. Getestet wurde auf die Anzahl der Datensätze und Wiederholungen. Die Besteinstellung ergab sich mit 300 Datensätzen und 400 Wiederholungen. Dabei ist zu beachten, dass 400 die Obergrenze darstellt. Allerdings liegt die Berechnungsdauer für diese Einstellung bereits bei zwei bis drei Stunden.
Bei der logistischen Regression liefert die jährliche Einstellung allgemein die besten Ergebnisse. Auch hier sind, wie beim linearen Modell, neben dem Referenzzeitpunkt April 2013 die Zeitpunkte April 2012 und 2011 für die Besteinstellung verantwortlich. Interessant ist hier auch die Frage, ob ein allgemeines $\boldsymbol{\beta}$ oder ein speziell für jeden Weg ermitteltes $\boldsymbol{\beta}$ optimal ist. Die Tatsache, dass keinerlei Einstellung mit einem wegspezifischen $\boldsymbol{\beta}$ besser als irgendeine Einstellung mit einem allgemeinen $\boldsymbol{\beta}$ ist, spricht wohl für sich.
Auch beim Conditional Logit liefern ein allgemeines $\boldsymbol{\beta}$ mit jährlichen Daten die besten Ergebnisse. Neben dem Referenzzeitpunkt April 2013 sind die weiteren verwendeten Zeitpunkte hierbei April 2012 - 2007.

Für die Vormodelle dient folgende Überlegung als Einstellungsgrundlage. Neben dem zu schätzenden Zeitpunkt werden drei Zeitpunkte benötigt: Der Referenzzeitpunkt, auf den das angepasste Modell angewendet wird und die beiden Zeitpunkte der Vergangenheit, auf denen die Anpassung des Modells stattfindet. Für eine Entscheidung seien die Optimaleinstellungen der entsprechenden Hauptmodelle betrachtet. Die logistische Regression zeigte zwar für die jährliche Einstellung die beste Werte, allerdings lieferte das Heranziehen von Daten im monatlichen Abstand fast genauso gute Ergebnisse. Den Ausschlag gibt das naive Modell. Bei diesem zeigt die Verwendung weit zurückliegender Zeitpunkte eine deutliche Verschlechterung der Vorhersagekraft. Demzufolge wird mit dem zu schätzenden Monat April 2014, dem Referenzzeitpunkt April 2013 und den Vergangenheitszeitpunkten Februar und März 2013 für alle Modelle dieselben Zeitpunkte gewählt. Für keines der Modelle ist dies eine Optimaleinstellung, aber die erzeugten

Werte sind immerhin bezüglich derselben Zeitpunkte vergleichbar. Für das Neuronale Netz zeigte sich, dass drei innere Schichten mit je $p = |\boldsymbol{\beta}|$ Knoten optimal sind. Die Anzahl der Knoten ist willkürlich nach der Anzahl der eingegebenen Parameter gesetzt. Die geschätzten Outputs des linearen Modells liegen im Intervall $[0, 1]$ und die des Neuronalen Netzes im Intervall $[-1, 1]$ und müssen mittels eines Entscheidungswertes zu 0 und 1 zugeordnet werden. In 0.01-Schritten wurde dieser Wert auf seine Genauigkeit hin überprüft und diejenige Einstellung gewählt, welche die höchste Genauigkeit liefert. Der Entscheidungswert beim linearen Modell liegt bei 0.51, der des Neuronalen Netzes bei -1, womit alle eingegebenen Routen existieren werden.

7.3 Auswertung zu den Fehlermaßen der Hauptmodelle

In den folgenden Abschnitten wird eine Auswertung für jedes Fehlermaß vorgenommen. Auf den gezeigten Abbildungen existieren für jeden Wert zwei Abbildungen. Diese entstehen durch das Entfernen von Ausreißern. Jedes Modell produziert mehr oder weniger stark abweichende Ergebnisse. Einige davon weichen derart stark ab, dass sie die Werte der Fehlermaße verwässern. So kann ein eigentlich gutes Modell aufgrund weniger ungünstiger Prognoseergebnisse als schlecht angesehen werden. Um dies zu vermeiden, wird jeder Wert zweimal betrachtet. Einmal mit Ausreißern und einmal ohne Ausreißer. Die Ausreißer bestehen aus allen Prognosewerten deren Differenz mit dem wahren Beobachtungswert außerhalb des einfachen Whiskerintervalls der Boxplots liegt. Die Werte mit Ausreißer werden auch **unbereinigt** und die Werte ohne Ausreißer **bereinigt** genannt.
Zudem wurden alle Werte auch in verschiedenen Passagiergrößenklassen untersucht, um Trends zu erkennen. Je nach wirklicher Anzahl der Passagiere auf einer beobachteten Route, wird diese einer bestimmten Klasse zugeordnet:

- Klasse 1: 1 - 251 Passagiere
- Klasse 2: 252 - 631 Passagiere
- Klasse 3: 632 - 1 585 Passagiere
- Klasse 4: 1 586 - 3 981 Passagiere
- Klasse 5: 3 982 - 10 000 Passagiere
- Klasse 6: 10 001 - 25 118 Passagiere
- Klasse 7: > 25 118 Passagiere.

Die Zahlen entsprechen den gerundeten Werten von $10^{2.4}$, $10^{2.8}$, $10^{3.2}$, $10^{3.6}$, $10^{4.0}$, $10^{4.4}$.
Wenn keine weiteren Bemerkungen gemacht werden, bilden die in den nachfolgenden Abschnitten gezeigten Abbildungen immer die Werte der Klasse 1 ab.

7.3.1 Boxplots

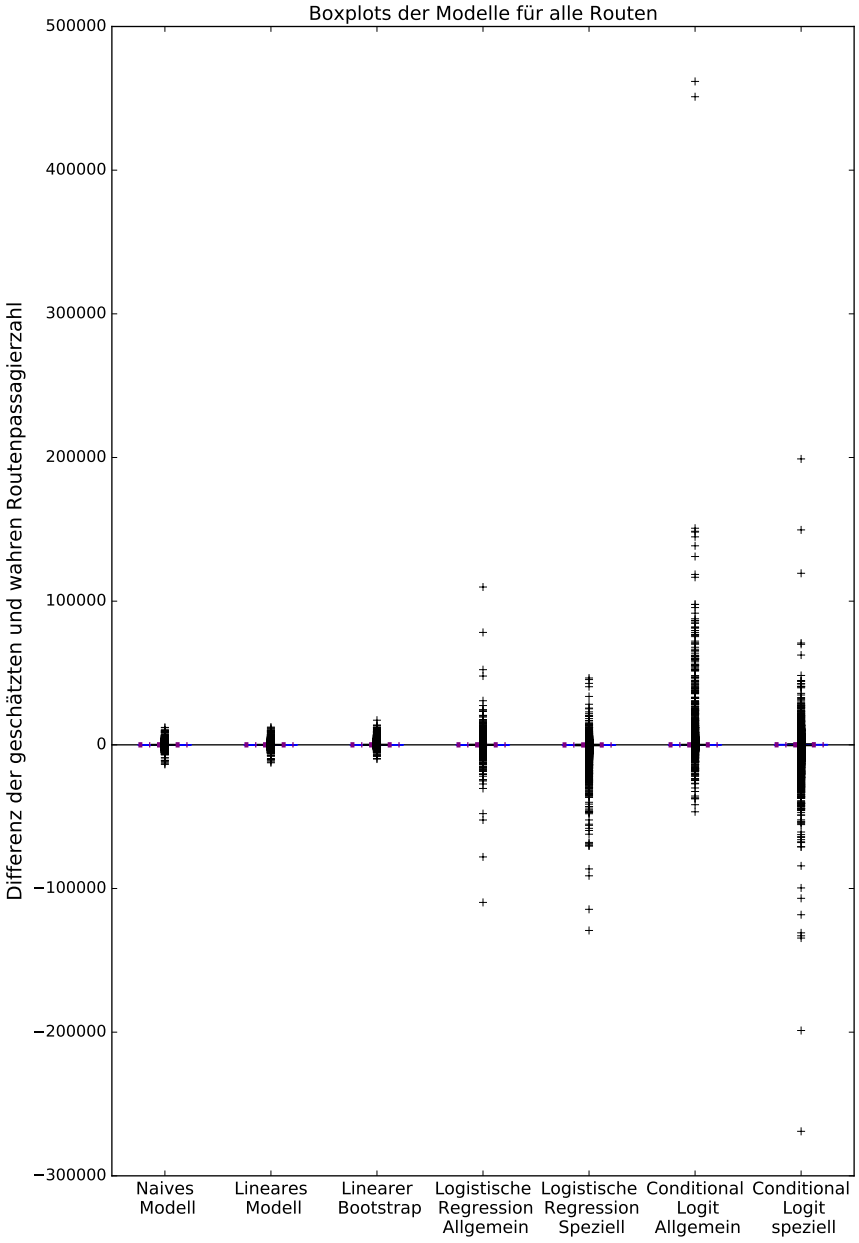


Abbildung 7.1: Boxplots für alle Routen des zu schätzenden Zeitpunktes April 2014. Gezeigt werden die Boxplots der Werte der Differenz aus der geschätzten und tatsächlichen Passagierzahl einer Route des zu schätzenden Zeitpunktes. Die pinke gestrichelte Linie gibt den Median der Werte an, das obere und untere Ende des blauen Kastens das 0.75 und 0.25 Quantil, das obere und untere Ende des gestrichelten Linien sind die Whisker, welche die Werte des 1.5-fachen der Quantile ergeben. Die Kreuze über und unter einem Boxplot stellen die Ausreißer dar, jene Werte, welche zu groß sind. Je ein Boxplot für jedes Modell.

Die in Abbildung 7.1 dargestellten Boxplots zeigen die Boxplots mit Ausreißern. Es ist zu erkennen, dass die Anzahl der Ausreißer nach oben und unten gleichmäßig verteilt erscheint. Auffällig sind die starken Ausreißer bei der logistischen Regression und vor allem bei dem Conditional Logit.

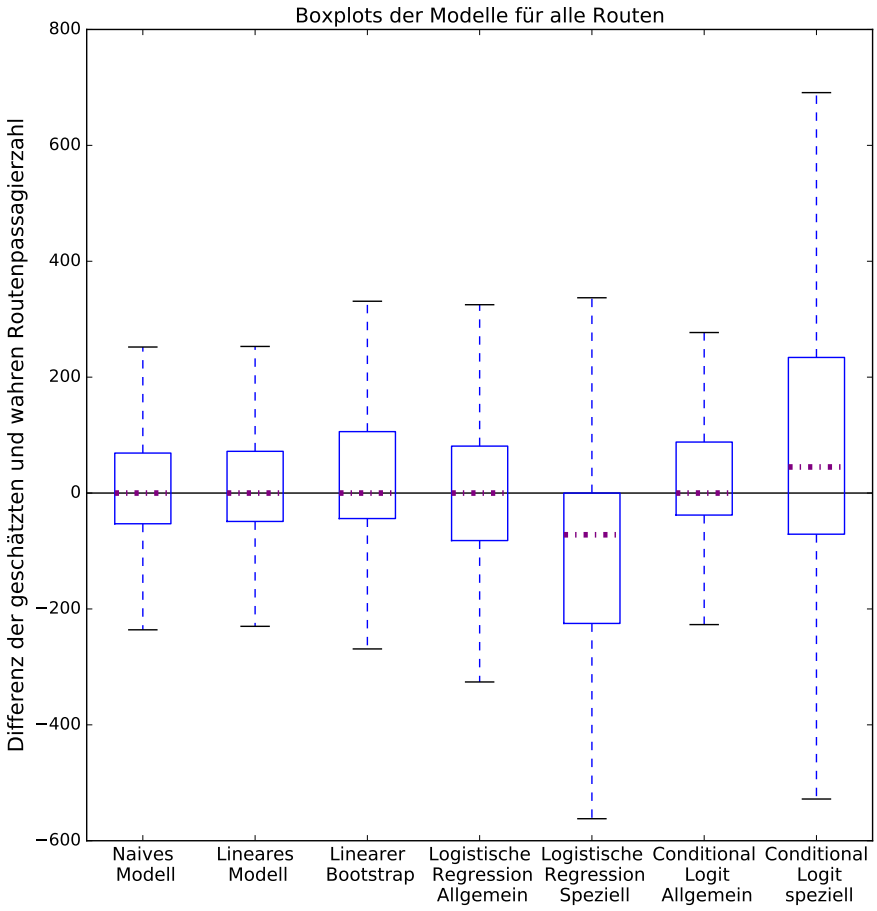


Abbildung 7.2: Boxplots für alle Routen des zu schätzenden Zeitpunktes April 2014. Gezeigt werden die Boxplots der Werte der Differenz aus der geschätzten und tatsächlichen Passagierzahl einer Route des zu schätzenden Zeitpunktes. Die pinke gestrichelte Linie gibt den Median der Werte an, das obere und untere Ende des blauen Kastens das 0.75 und 0.25 Quantil, das obere und untere Ende des gestrichelten Linien sind die Whisker, welche die Werte des 1.5-fachen der Quantile ergeben. Je ein Boxplot für jedes Modell.

Aufgeteilt in die Klassen ergibt sich eine ähnliche Auffälligkeit bezüglich der Ausreißer. Hierbei ist zu bemerken, dass die Ausreißer in Klasse 1 fast vollständig im positiven Bereich angesiedelt sind und in Klasse 2 fast vollständig im negativen Bereich. In den Klassen dazwischen dreht sich dieses Verhältnis entsprechend der Klassenzahl. In Klasse 6 erscheint das Verhältnis ausgewogen.
Die Werte für eine Route entstehen aus der Differenz aus prognostizierter Passagier-

zahl und der wahren beobachteten Passagierzahl. Die Klassenbilder lassen also den Schluss zu, dass Routen mit wenigen Passagieren tendenziell überschätzt und Routen mit großer Passagierzahl tendenziell unterschätzt werden. „Unterschätzt“ meint, dass weniger Passagiere bestimmt werden, als eigentlich auf der Route unterwegs sind. Das analoge Gegenteil gilt für „überschätzt“.

Dass die eigentlichen Boxplots nicht zu erkennen sind, lässt darauf schließen, dass sich der Großteil der Differenzen nahe dem Wert 0 bewegt. Dazu sei Abbildung 7.2 betrachtet. Auffällig ist, dass der Median der Differenzen genau auf der 0 liegt. Dies bedeutet, dass sich die 0 in der Mitte der sortierten Differenzenwerte befindet. Bei derartig vielen Routen und der Tatsache, dass die bei den meisten Modellen geschieht lässt den Schluss zu, das dies kein Zufall ist. Bei einer Auflistung der Differenzenwerte ergab sich, dass die Anzahl des Auftretens des Wertes 0 in einem vierstelligen Bereich liegt, wohingegen der Wert 1 in einem niedrigen zweistelligen Bereich liegt. Die Anzahlen anderer Differenzen werden kleiner, je größer ihr Wert ist.
Auffällig ist diesmal, dass der allgemeine Conditional Logit sehr gut im Vergleich mit den anderen Modellen abschneidet.

Aufgeteilt in die Klassen bestätigt sich der Trend des guten Conditional Logit. Liegen in den niedrigsten Klassen alle Modelle mit ihren Boxplots noch nah beieinander, so ist schon ab Klasse 3-4 eindeutig zu erkennen, dass die logistische Regression und der Conditional Logit stark streuen. Mit Ausnahme des allgemeinen Conditional Logits. Dieser besitzt ähnliche Werte wie das lineare oder das naive Modell und ist zeitweise sogar besser als diese, wenngleich er auch eindeutige Tendenzen zur Unterschätzung aufweist. Die Streuungen sind als extrem, aber vergleichsweise gering in ihrer Anzahl.
Die speziellen Varianten der logistischen Regression und des Conditional Logits zeigen sehr starke Streuungen, wobei der Boxplot des speziellen Conditional Logits vor allem im oberen Klassenbereich mit einer teilweise 20-fachen Größe alle anderen Boxplots überbietet. Der spezielle Conditional Logit ist bereits an dieser Stelle als anwendbares Modell auszuschließen. Auch die spezielle logistische Regression weist einen vergleichsweise großen Boxplot auf, zudem der Median nicht auf der 0 liegt. Er ist demzufolge ebenso wenig zu empfehlen, da es eindeutig bessere Alternativen gibt.
Die Tendenzen der Über- und Unterschätzung in den einzelnen Klassen bestätigen sich.

Allgemein lässt sich sagen, dass sowohl das naive Modell als auch das lineare Modell und der allgemeine Conditional Logit zu den favorisierten Modellen gezählt werden können. Der Boostrap ist in der Größe des Boxplots leicht unterlegen. Gegen das naive Modell spricht die Abhängigkeit der Daten vom gewählten Monat und gegen den Conditional Logit die starken Ausreißer, welche bei den anderen beiden Modellen nicht auftreten.

7.3.2 ME - Mittlerer Fehler

Abbildung 7.3 zeigt den mittleren Fehler der Modelle und in Tabelle 7.5 sind die genauen Werte aufgeführt. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.4.

Modell	ME mit Ausreißer Angabe in	ME ohne Ausreißer Angabe in Passagieren
Naives Modell	14.6018	4.2380
Lineares Modell	22.6472	5.5695
Linearer Bootstrap	73.866	16.8512
Allgemeine Logistische Regression	58.0275	-2.2697
Spezielle Logistische Regression	-212.6722	-91.3098
Allgemeiner Conditional Logit	124.4780	15.4264
Spezieller Conditional Logit	-27.1059	66.8263

Tabelle 7.5: Tabelle des ME für die Hauptmodelle, mit und ohne Ausreißern

Es ist gut zu erkennen, dass sowohl das naive Modell als auch das lineare Modell sehr gute Werte nahe 0 aufweisen, sowohl bei den bereinigten als auch bei den unbereinigten Werten.
Nach der Entfernung der Ausreißer zeigen auch der lineare Bootstrap und die allgemeine logistische Regression recht gute Werte nahe 0. Im Allgemeinen bestätigt sich die Aussage des Nullmedians. Auch die Ausreißer scheinen sich nach oben und unten die Waage zu halten. Vergleichsweise extreme Werte liefern die letzten drei Modelle.
Auffälligkeiten zeigt der spezielle Conditional Logit. Vor der Bereinigung hatte er negative Werte und danach positive. Alle anderen Modelle haben lediglich ihren Trend in abgeschwächter Weise bestätigt. Nur beim Conditional Logit ist zu sehen, dass die Ausreißer eher ins Negative tendieren, also zu einer Überschätzung neigen.

Über die Passagiergrößenklassen hinweg bestätigen sich obige Aussagen, vor allem der spezielle Conditional Logit weist im hohen Bereich viele negative Werte auf. Auch die Tendenz zu den Ausreißern vom vorherigen Abschnitt spiegelt sich in den Werten wider.

Nach Bild 7.3 sind das naive und das lineare Modell den anderen vorzuziehen. Der noch im vorherigen Abschnitt als gut eingestufte allgemeine Conditional Logit verliert an Attraktivität aufgrund seiner bereits erwähnten Ausreißer.

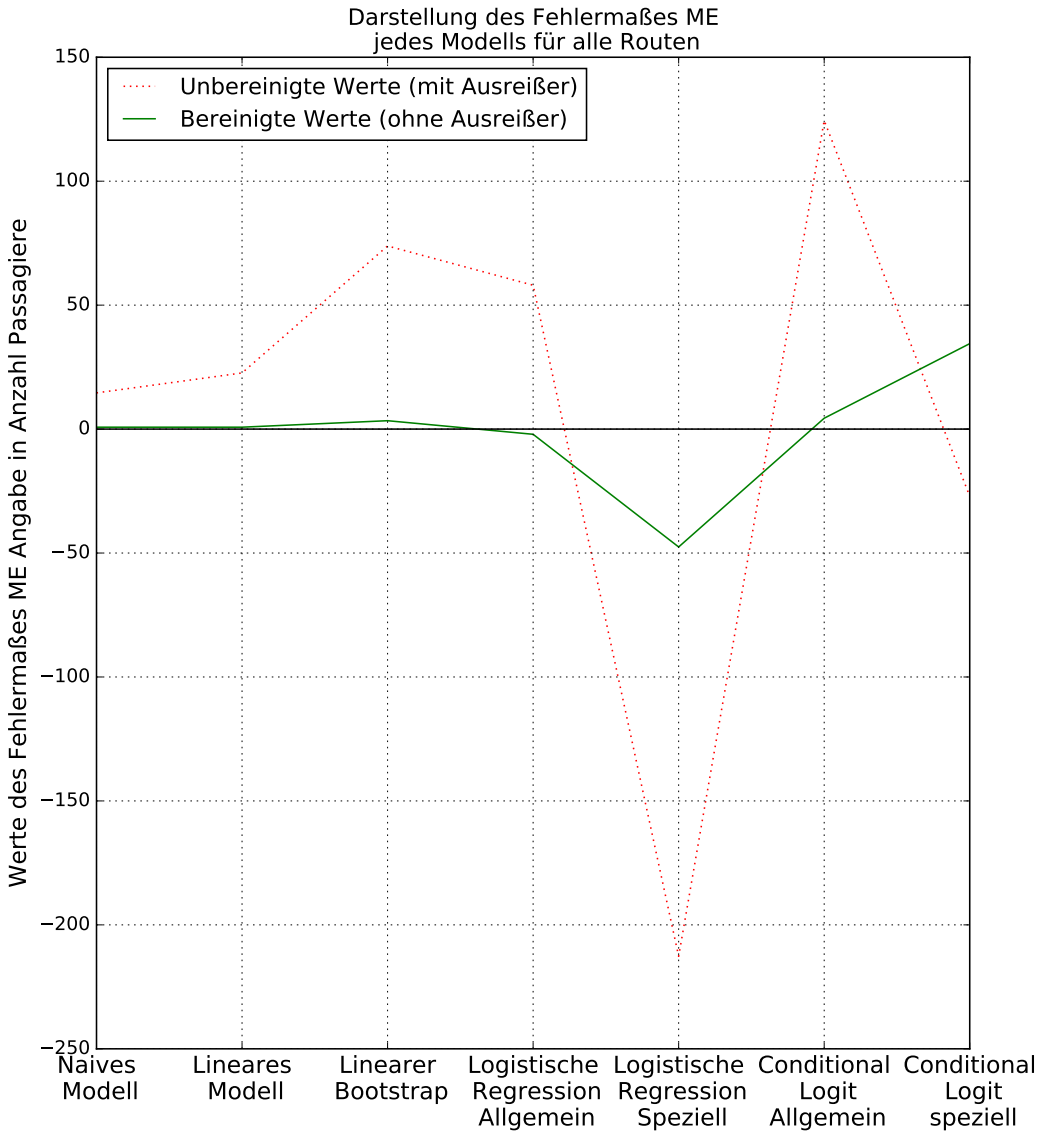


Abbildung 7.3: Darstellung des Fehlermaßes ME für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

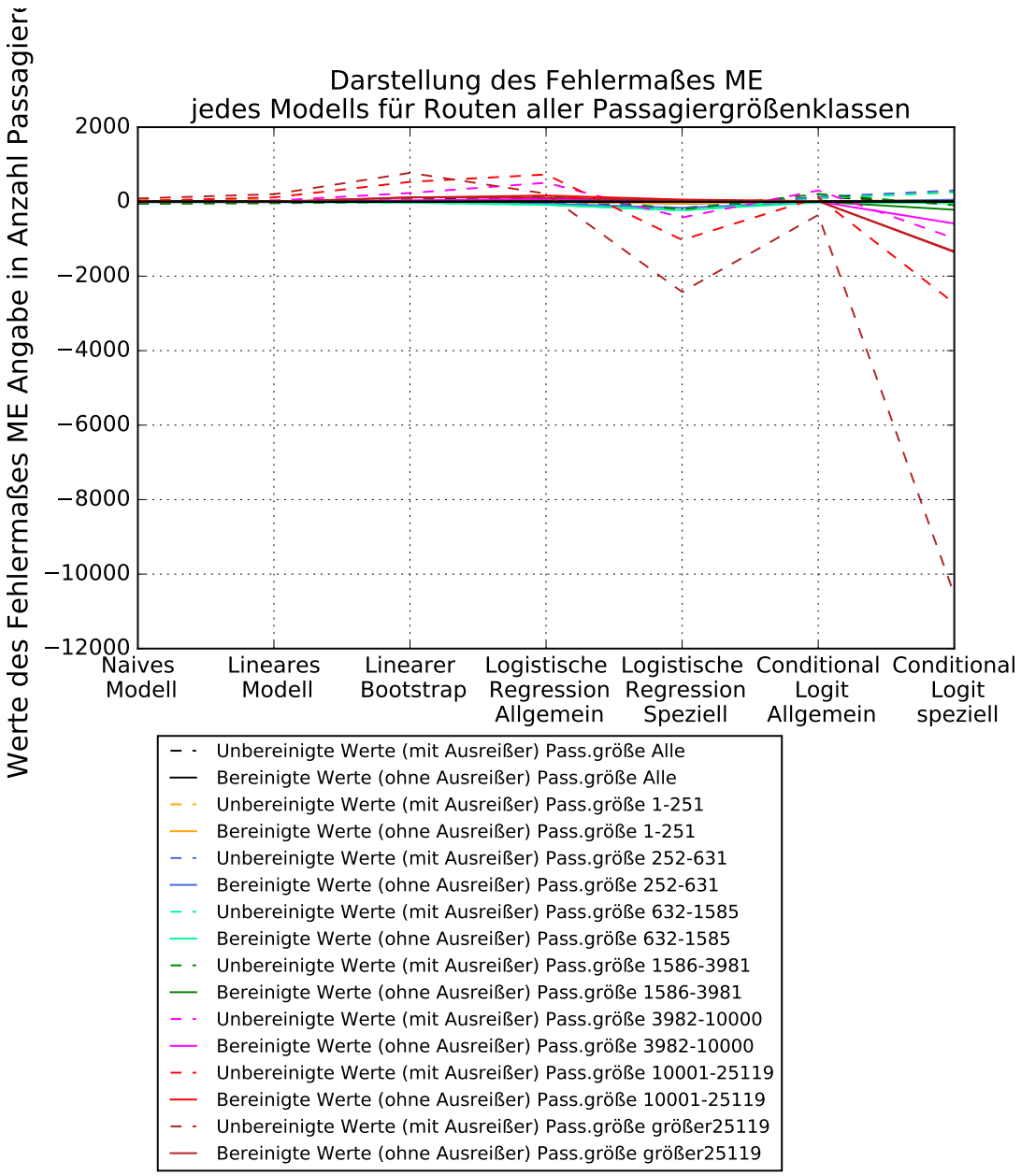


Abbildung 7.4: Darstellung des Fehlermaßes ME für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.3.3 MAE - Mittlerer absoluter Fehler

Abbildung 7.5 und Tabelle 7.6 zeigen den mittleren absoluten Fehler. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.6.

Bezüglich der Genauigkeit der Vorhersagen ist dieser Fehler eines der am besten zu vergleichenden Maße. Negative und positive Werte werden nicht gegeneinander aufgerechnet. Die Abweichungen sind absolut und besitzen dieselbe Maßeinheit wie die Prognosewerte. Die angezeigten Ausprägungen der bereinigten Variante sind in fast allen Passagiergrößenklassen ähnlich.

Modell	MAE mit Ausreißer Angabe in Passagieren	MAE ohne Ausreißer Angabe in Passagieren
Naives Modell	168.0202	60.3024
Lineares Modell	161.4667	60.0646
Linearer Bootstrap	217.9892	73.5901
Allgemeine Logistische Regression	286.6516	78.2598
Spezielle Logistische Regression	374.5039	126.4054
Allgemeiner Conditional Logit	289.1672	62.3129
Spezieller Conditional Logit	645.4860	161.6372

Tabelle 7.6: Tabelle des MAE für die Hauptmodelle, mit und ohne Ausreißern

Das angezeigte Bild hat in den unteren Passagiergrößenklassen sehr starke Ausprägungen bei der logistischen Regression und dem speziellen Conditional Logit. Der Conditional Logit besitzt vor allem im hohen Klassenbereich extreme Ausprägungen. Die Differenzenwerte bewegen sich bei der bereinigten Variante der favorisierten Modelle erstaunlicherweise auch in höheren Passagiergrößenklassen im Bereich von 70-150 Passagieren, was einen äußerst positiven Kritikpunkt darstellt. Immerhin bewegt sich der Großteil der Passagiere auf den Routen der hohen Passagiergrößenklassen, womit höhere Werte zu erwarten gewesen wären.

Mit diesem Bild bestätigt sich die Vermutung, die in den Boxplots gewonnen wurde, bestätigen. Das naive und das lineare Modell liefern ähnlich gute Werte. Der allgemeine Conditional Logit wäre ohne seine Ausreißer als ebenso gut einzustufen. Der lineare Bootstrap ist stets geringfügig schlechter als die ersten beiden Modelle.

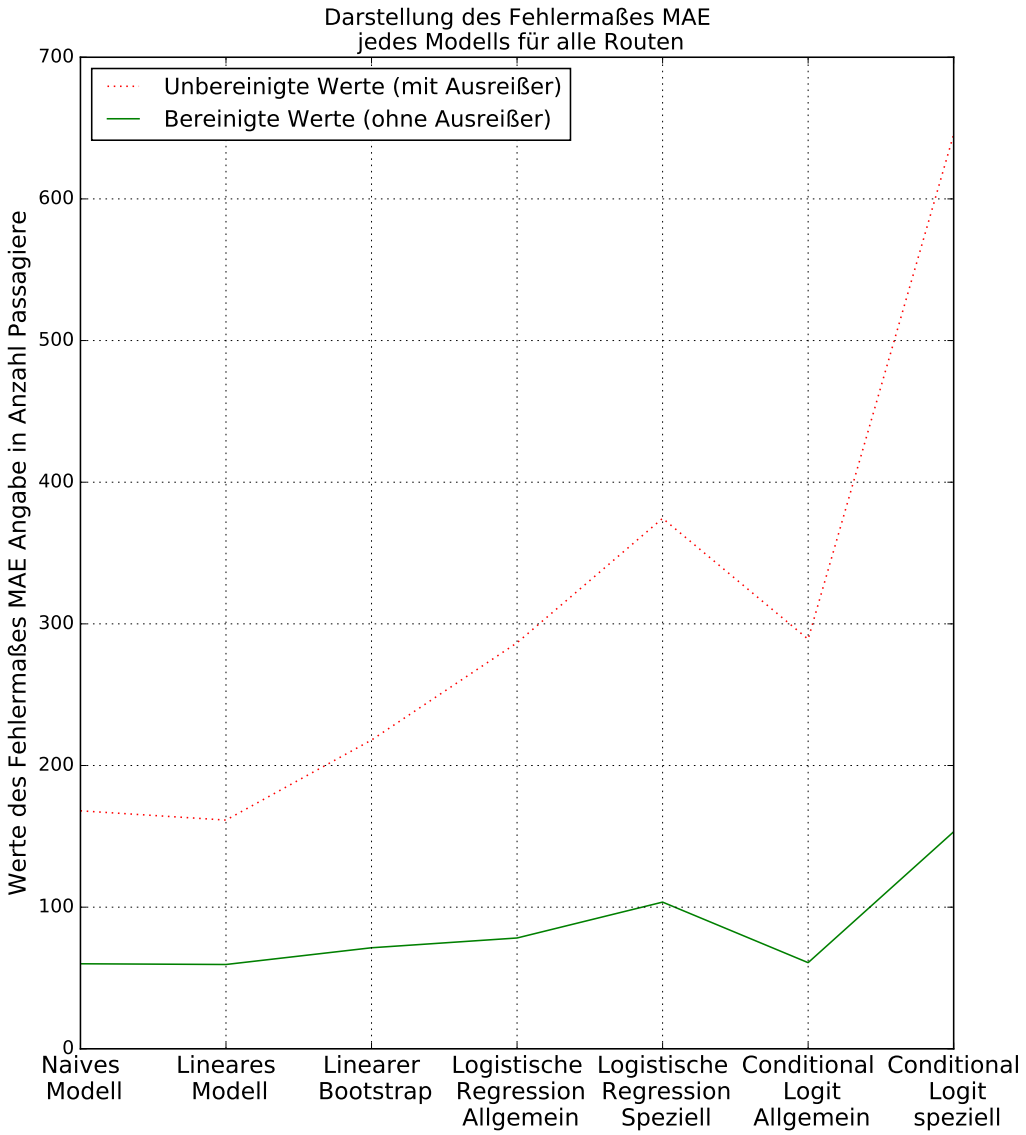


Abbildung 7.5: Darstellung des Fehlermaßes MAE für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

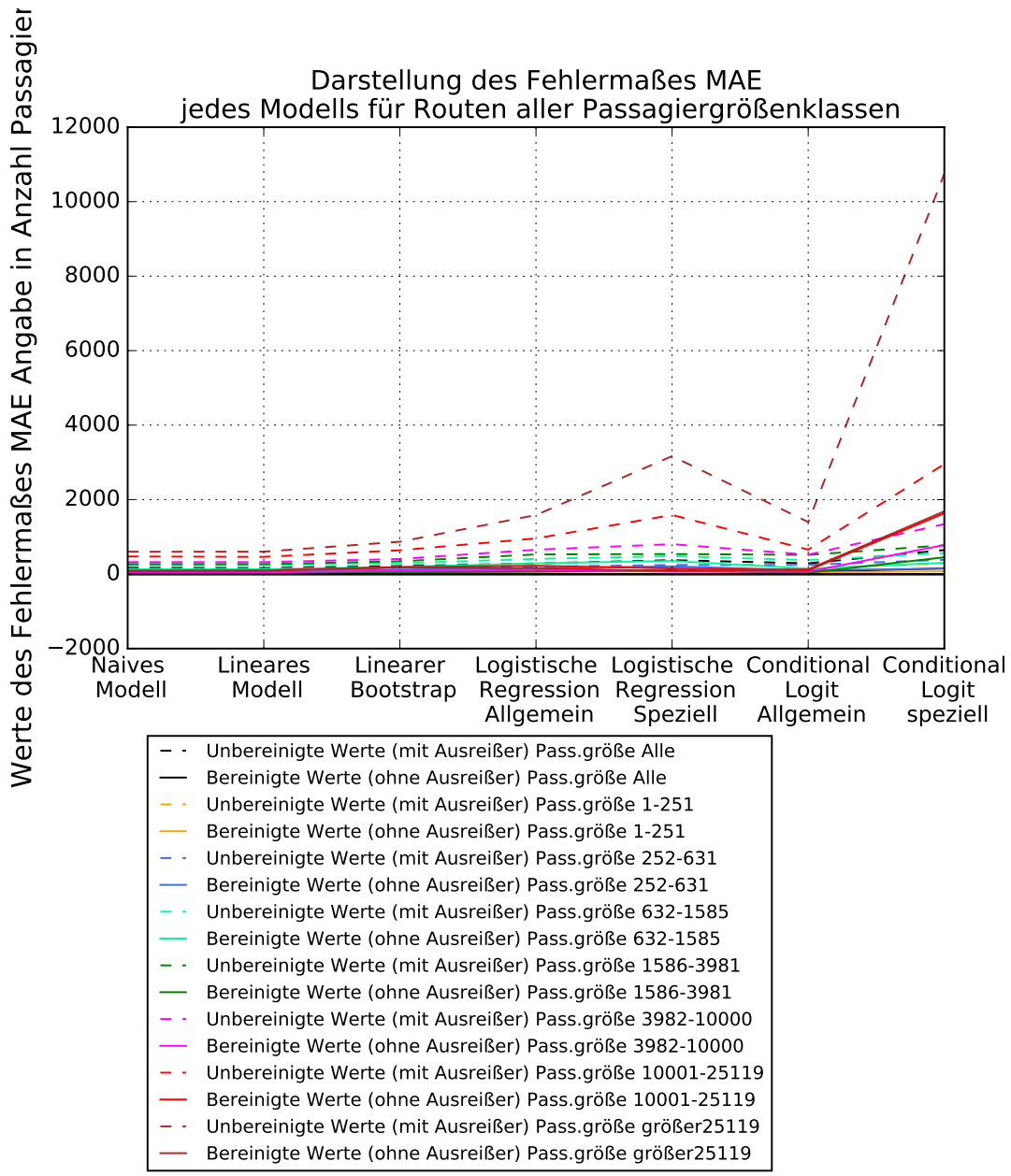


Abbildung 7.6: Darstellung des Fehlermaßes MAE für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.3.4 MPE - Mittlerer prozentualer Fehler

Abbildung 7.7 und Tabelle 7.7 zeigen den mittleren prozentualen Fehler. Die Abbildung entspricht dem Aussehen der Passagiergrößenklassen 1-3. Aufgrund des veränderten Aussehens ab Klasse 4 sei eine weitere Abbildung 7.8 eingefügt, welche symptomatisch für die oberen Klassen ist. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.9.

Modell	MPE mit Ausreißer Angabe in Prozent	MPE ohne Ausreißer Angabe in Prozent
Naives Modell	15.4602	6.2868
Lineares Modell	14.6674	5.3441
Linearer Bootstrap	19.7079	8.2588
Allgemeine Logistische Regression	6.3091	-1.0120
Spezielle Logistische Regression	-25.920	-32.7560
Allgemeiner Conditional Logit	49.0497	10.4611
Spezieller Conditional Logit	110.5277	46.9334

Tabelle 7.7: Tabelle des MPE für die Hauptmodelle, mit und ohne Ausreißern

Die Tendenzen der Unter- und Überschätzung in den Passagierklassen aus den Boxplots wird erneut bestätigt. Wenn auch die bereinigte allgemeine logistische Regression nach Abbildung 7.7 die besten Werte aufweist, so ist dieser Schluss bei genauerer Betrachtung nicht zulässig. Wie auf Abbildung 7.8 zu sehen ist, liefert dieses Modell bei unbereinigten Daten sogar die schlechtesten Werte. Allerdings gleichen sich die positiven und negativen Werte insgesamt aus, sodass diese falsche Güte zu erkennen ist.

Die speziellen Varianten liefern erneut keine guten Werte, wohingegen das naive und das lineare Modell wieder gut abschneiden. Der allgemeine Conditional Logit ist in den meisten Klassen besser als der lineare Bootstrap, wobei seine Ausreißer im unbereinigten Fall erneut zu teils sehr hohen Abweichungen führen.

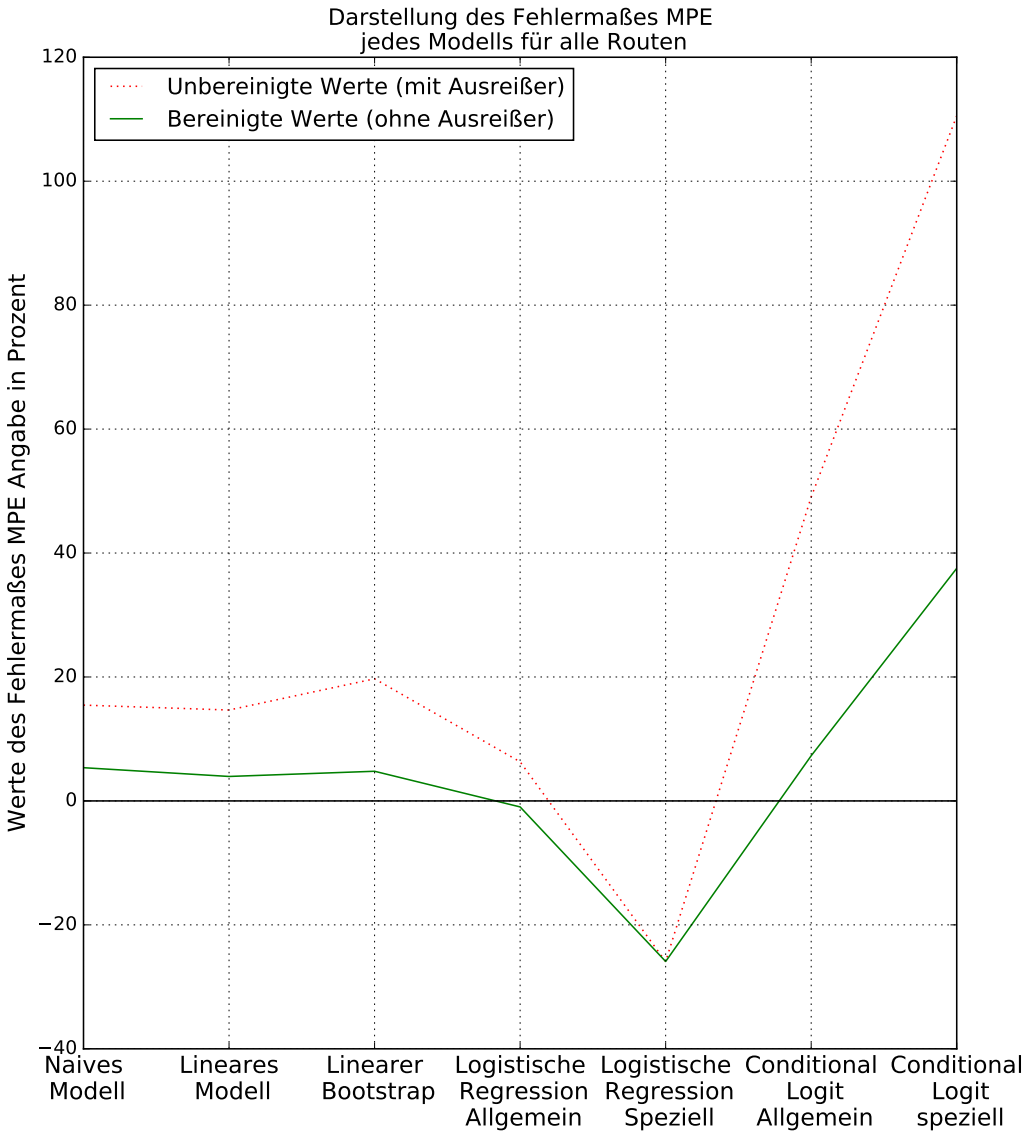


Abbildung 7.7: Darstellung des Fehlermaßes MPE für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

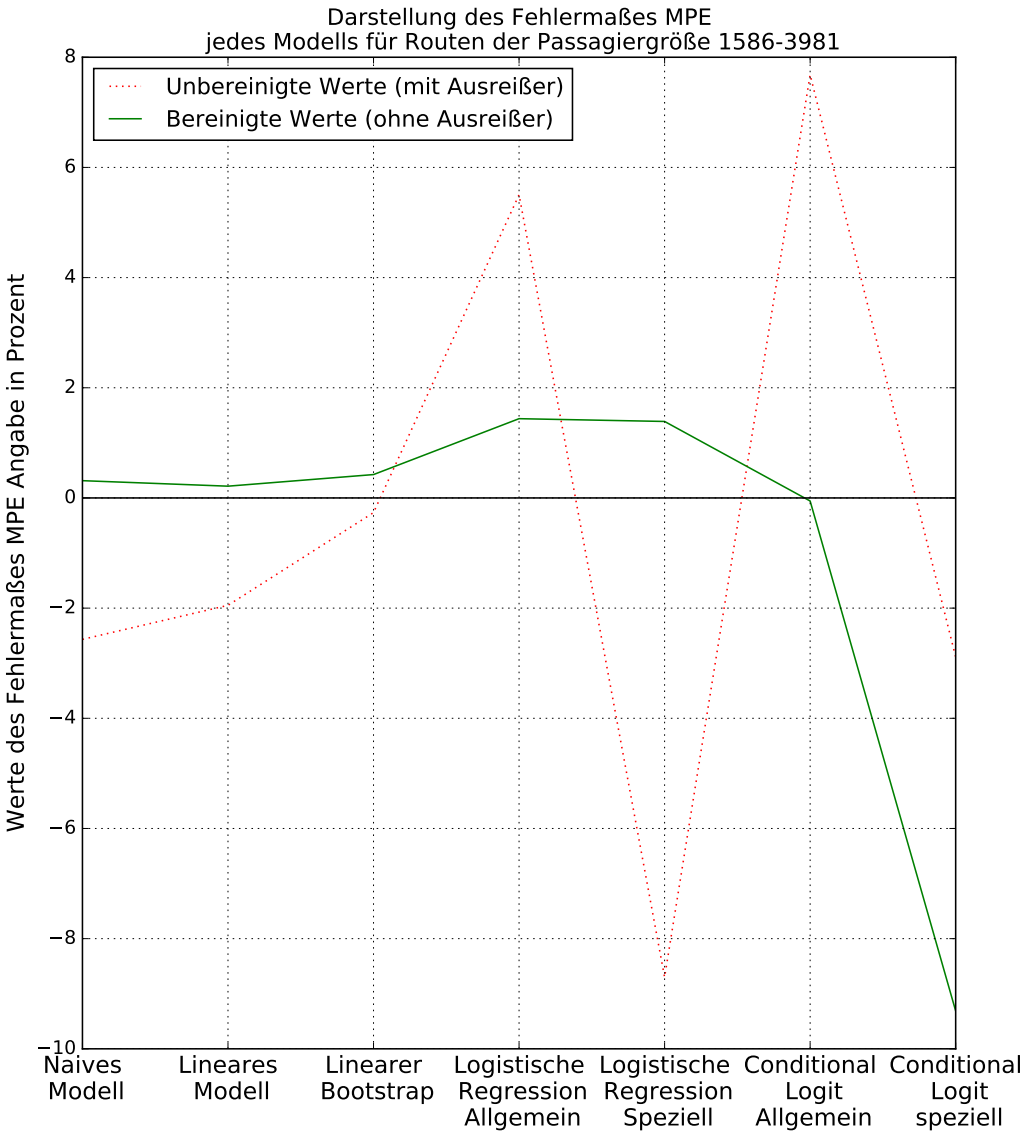


Abbildung 7.8: Darstellung des Fehlermaßes MPE jedes Modelles für die Passagiergrößenklasse 4: 1 586-3 981 Passagiere. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

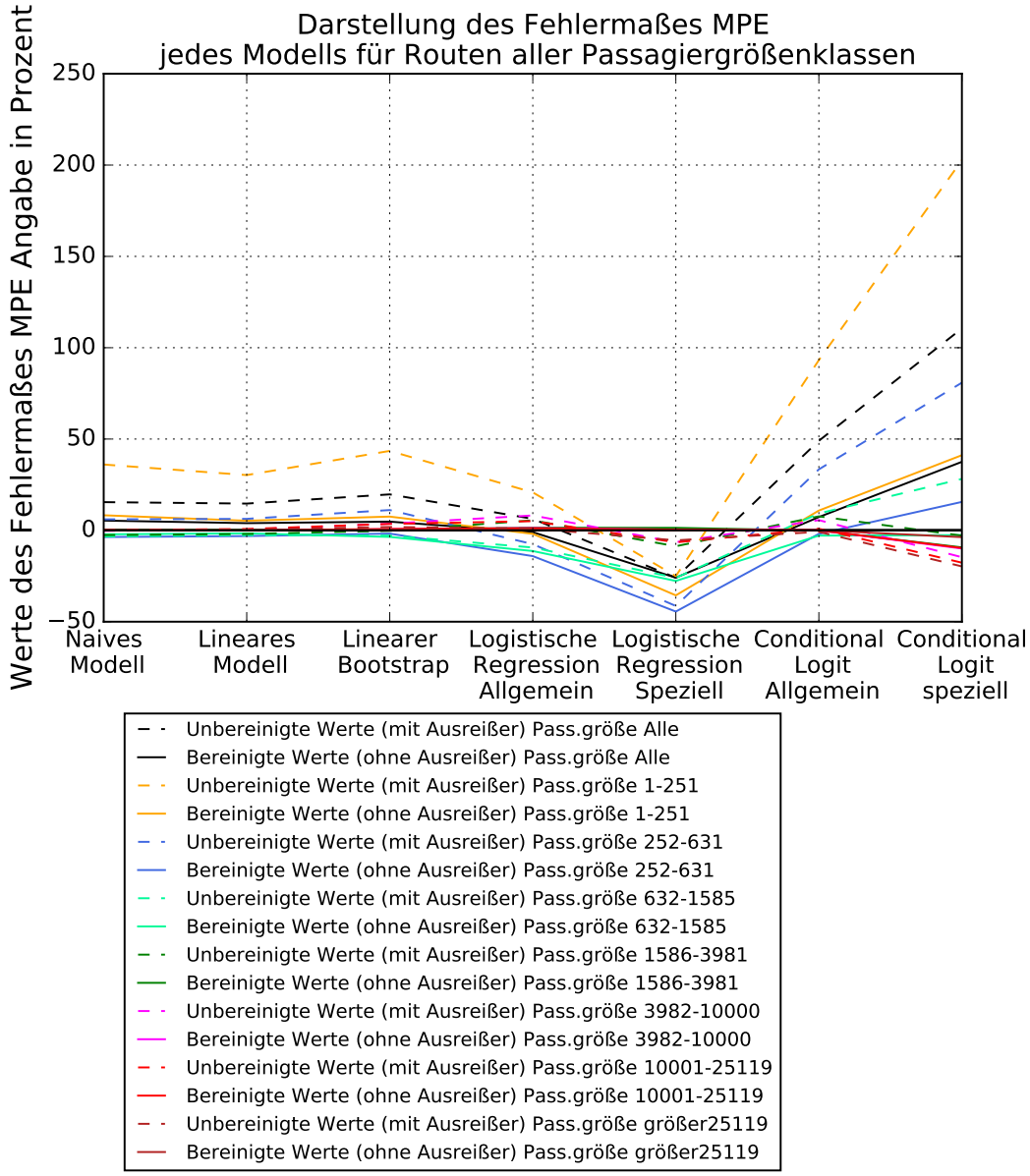


Abbildung 7.9: Darstellung des Fehlermaßes MPE für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.3.5 MAPE - Mittlerer absoluter prozentualer Fehler

Abbildung 7.10 und Tabelle 7.8 zeigen den mittleren absoluten prozentualen Fehler der Modelle. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.9. Aufgrund der Wichtigkeit der Werte ist eine erweiterte Tabelle 7.9 angegeben. Bezüglich der Genauigkeit der Vorhersagen ist dieser Fehler eines der am besten zu vergleichenden Maße. Negative und positive Werte werden nicht gegeneinander aufgerechnet. Zu beachten ist, dass bei zwei vorhergesagten Passagieren anstelle von einem bereits 100% Abweichung vorherrscht. Damit sind teils hervorragende Werte in den hohen Passagiergrößenklassen von 3-5% strenger zu bewerten als 30% Abweichung in den niedrigen Klassen.

Modell	MAPE mit Ausreißer Angabe in Prozent	MAPE ohne Ausreißer Angabe in Prozent
Naives Modell	37.0727	24.1789
Lineares Modell	39.0800	26.2467
Linearer Bootstrap	44.8931	28.9170
Allgemeine Logistische Regression	48.2157	32.7562
Spezielle Logistische Regression	54.5337	44.4715
Allgemeiner Conditional Logit	69.4433	25.8758
Spezieller Conditional Logit	130.9070	64.2446

Tabelle 7.8: Tabelle des MAPE für die Hauptmodelle, mit und ohne Ausreißern

Wie in Abbildung 7.11 zu sehen ist, entsprechen die Abbildungen aller Passagiergrößenklassen in ihren Ausprägungen in etwa denen von Abbildung 7.10. Im Allgemeinen lässt sich sagen, dass sich alle bisher getroffenen Vermutungen zur Güte der verschiedenen Modelle bestätigen. Das naive Modell, das lineare Modell und das Modell des allgemeinen Conditional Logit liefern die genauesten Werte wenn die Ausreißer herausgerechnet werden. Sie erfüllen im Mittel über alle Routen sogar das vom DLR vorgegebene Ziel von 25% Abweichung oder besser. Es ist dennoch erstaunlich, zu wie viel Abweichung die Ausreißer des allgemeinen Conditional Logit führen. Die speziellen Modelle liefern die schlechtesten Werte.
Die Routen der beiden kleinsten Klassen führen zu den vergleichsweise hohen Vorhersagefehlern. So ist beträgt die Abweichung in Klasse 1 fast 40% und in Klasse ungefähr 27% in den guten bereinigten Modellen. Bereits ab Klasse 3 liegen diese Werte bei nur noch 15%. Im Allgemeinen lässt sich damit behaupten, dass die guten Modelle vor allem die Routen mit vielen Passagieren recht genau vorhersagen können, wohingegen Routen mit wenigen Passagieren Schwierigkeiten bereiten. Eine gesonderte Berechnung für große und kleine Routen könnte helfen, die Genauigkeit zu verbessern.
Erstaunlich ist, dass sowohl das lineare Modell als auch der allgemeine Conditional Logit zu guten Prognosewerten führen. Und das, obwohl das lineare Modell lediglich einen Anpassungsparameter (die Passagierzahl) besitzt und der allgemeine Conditio-

nal Logit 22. Dies lässt darauf schließen, dass für den Conditional Logit viele irrelevante Werte verwendet werden. Mittels Analysemethoden ist zu klären, welche Parameter zu entfernen sind.

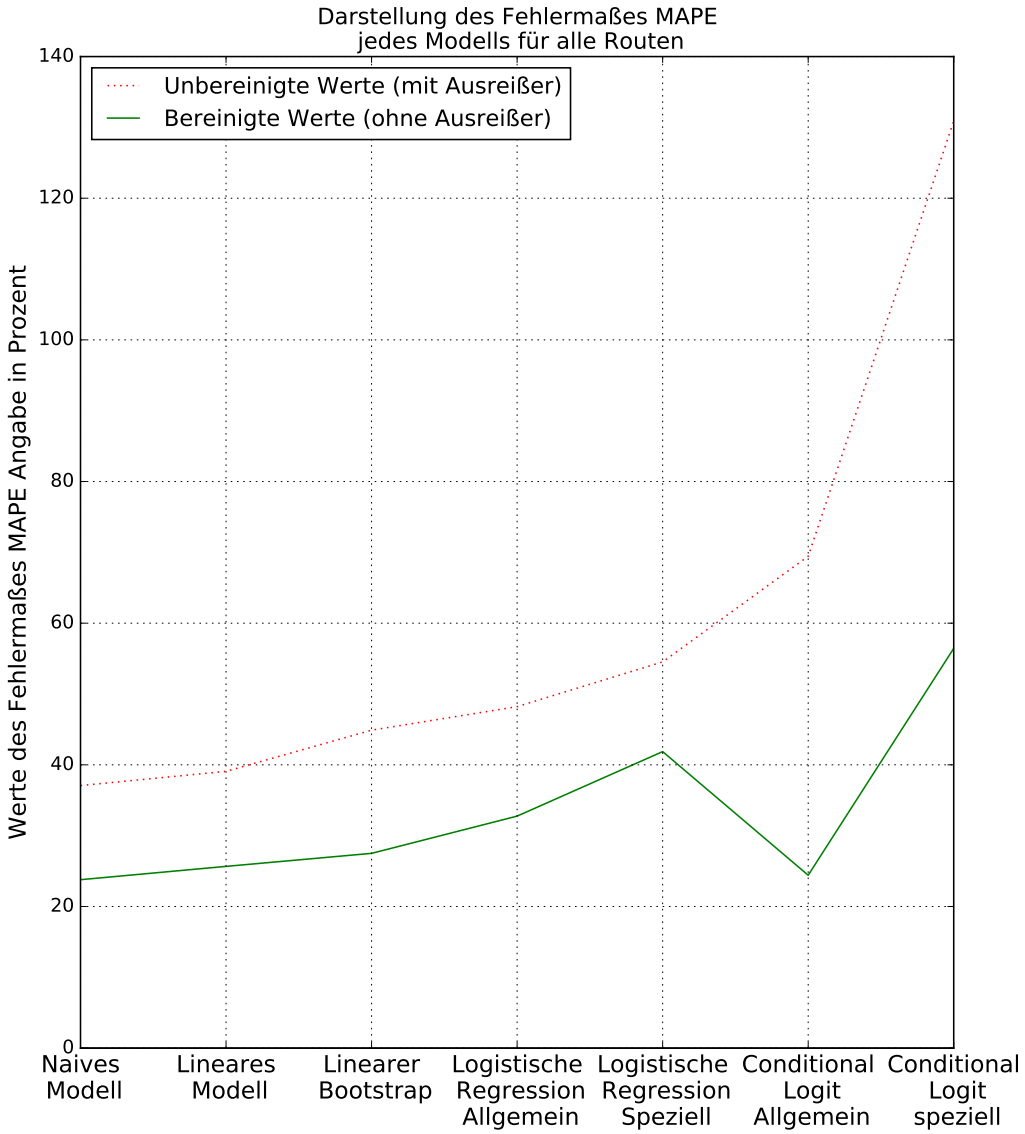


Abbildung 7.10: Darstellung des Fehlermaßes MAPE für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

Modell	Alle	Klasse 1	Klasse 2	Klasse 3	Klasse 4	Klasse 5	Klasse 6	Klasse 7
Mit Ausreißer								
Naives Modell	37.07	56.32	36.00	24.43	11.35	5.29	3.13	1.50
Lineares Modell	39.08	56.38	37.36	25.05	11.41	5.28	3.08	1.50
Linearer Bootstrap	44.89	69.87	43.72	30.26	14.49	6.65	4.28	2.12
Allgemeine Logistische Regression	48.21	64.81	50.53	42.05	22.23	10.72	6.44	3.54
Spezielle Logistische Regression	54.53	70.10	62.07	50.51	22.84	12.85	10.33	7.26
Allgemeiner Conditional Logit	69.44	110.26	62.66	39.39	21.26	9.17	4.25	3.45
Spezieller Conditional Logit	130.90	213.88	100.64	59.77	32.32	21.43	19.21	20.24
Ohne Ausreißer								
Naives Modell	24.17	34.21	25.80	16.07	2.16	0.88	0.64	0.20
Lineares Modell	26.24	36.10	27.13	15.74	1.77	0.87	0.73	0.28
Linearer Bootstrap	28.91	40.73	31.41	22.52	2.71	2.09	1.92	0.85
Allgemeine Logistische Regression	32.75	44.27	42.52	34.62	4.01	2.91	2.37	0.61
Spezielle Logistische Regression	44.47	55.77	57.08	43.82	4.27	1.09	0.98	0.20
Allgemeiner Conditional Logit	25.87	36.23	27.37	18.10	1.81	1.36	0.93	0.44
Spezieller Conditional Logit	64.24	93.47	46.82	32.26	21.32	17.88	18.15	18.39

Tabelle 7.9: Vollständige Tabelle des MAPE für die Hauptmodelle, mit und ohne Ausreißer für alle Passagiergrößenklassen

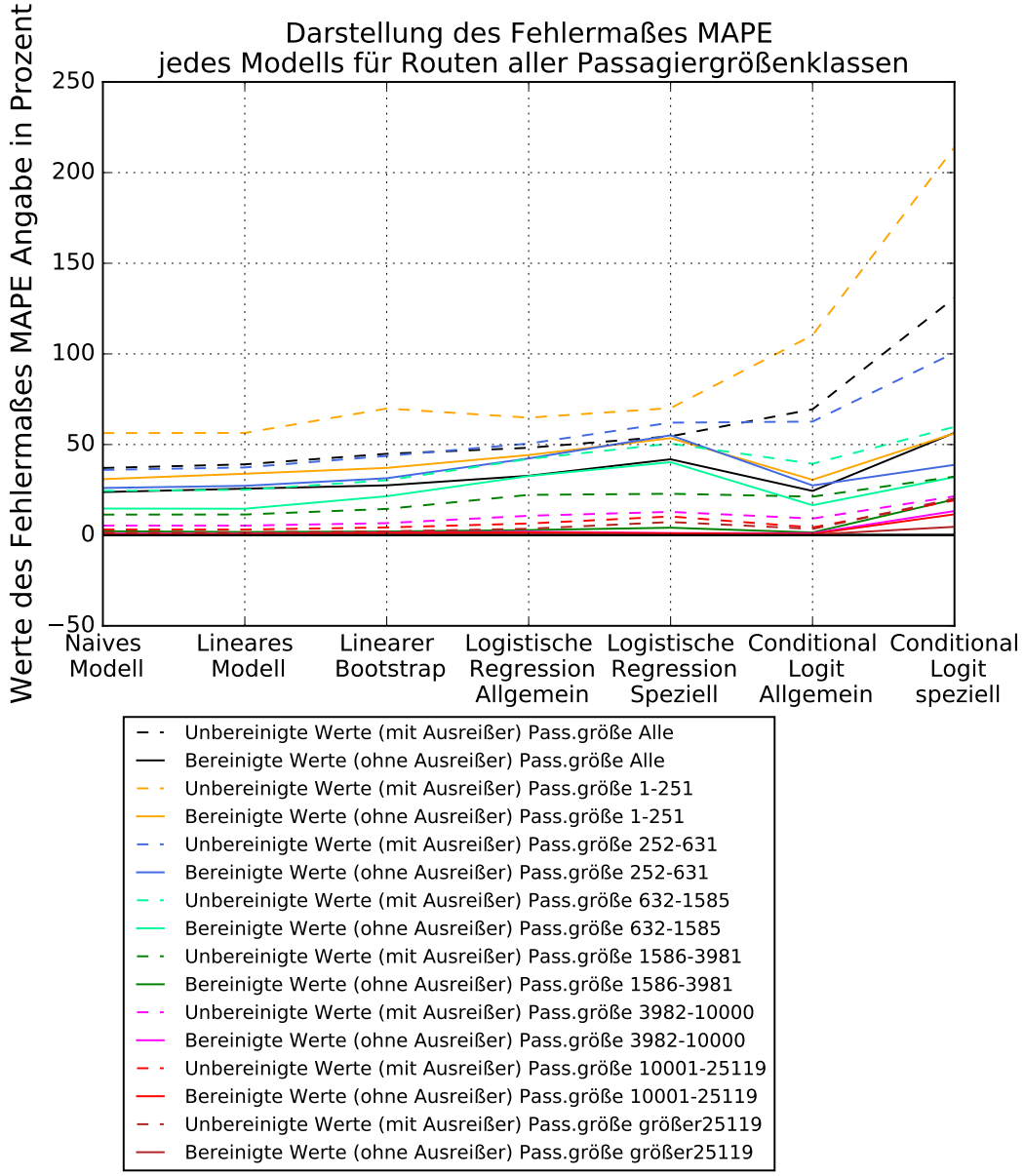


Abbildung 7.11: Darstellung des Fehlermaßes MAPE für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.3.6 MdAPE - Median des absoluten prozentualen Fehlers

Abbildung 7.12 und Tabelle 7.10 zeigen den mittleren absoluten prozentualen Fehler der Modelle. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.14.

Modell	MdAPE mit Ausreißer Angabe in Prozent	MdAPE ohne Ausreißer Angabe in Prozent
Naives Modell	16.9856	12.4573
Lineares Modell	18.8640	14.1509
Linearer Bootstrap	20.2884	14.1975
Allgemeine Logistische Regression	26.0593	17.3277
Spezielle Logistische Regression	47.5935	40.9448
Allgemeiner Conditional Logit	18.3098	12.6819
Spezieller Conditional Logit	37.7410	32.3033

Tabelle 7.10: Tabelle des MdAPE für die Hauptmodelle, mit und ohne Ausreißern

Abbildung 7.12 bestätigt die Vermutung des vorherigen Abschnittes 7.3.5, dass die kleinen Passagiergrößenklassen 1-3 für starke Abweichungen im durchschnittlichen Medianwert führen.
Denn wie in 7.13 zu sehen ist, besitzen alle anderen Klassen einen Median, der bei knapp 1 oder 0% liegt. Dies bedeutet, dass vor allem viele große Klassen fast exakt vorhergesagt werden, während die Vorhersagen in den kleinen Klassen vergleichsweise stark von den realen Werten abweichen.

Auffällig ist, dass in den Klassen 4-7 alle Modelle außer dem speziellen Conditional Logit einen Median von ungefähr 1% für die bereinigten Werte besitzen.

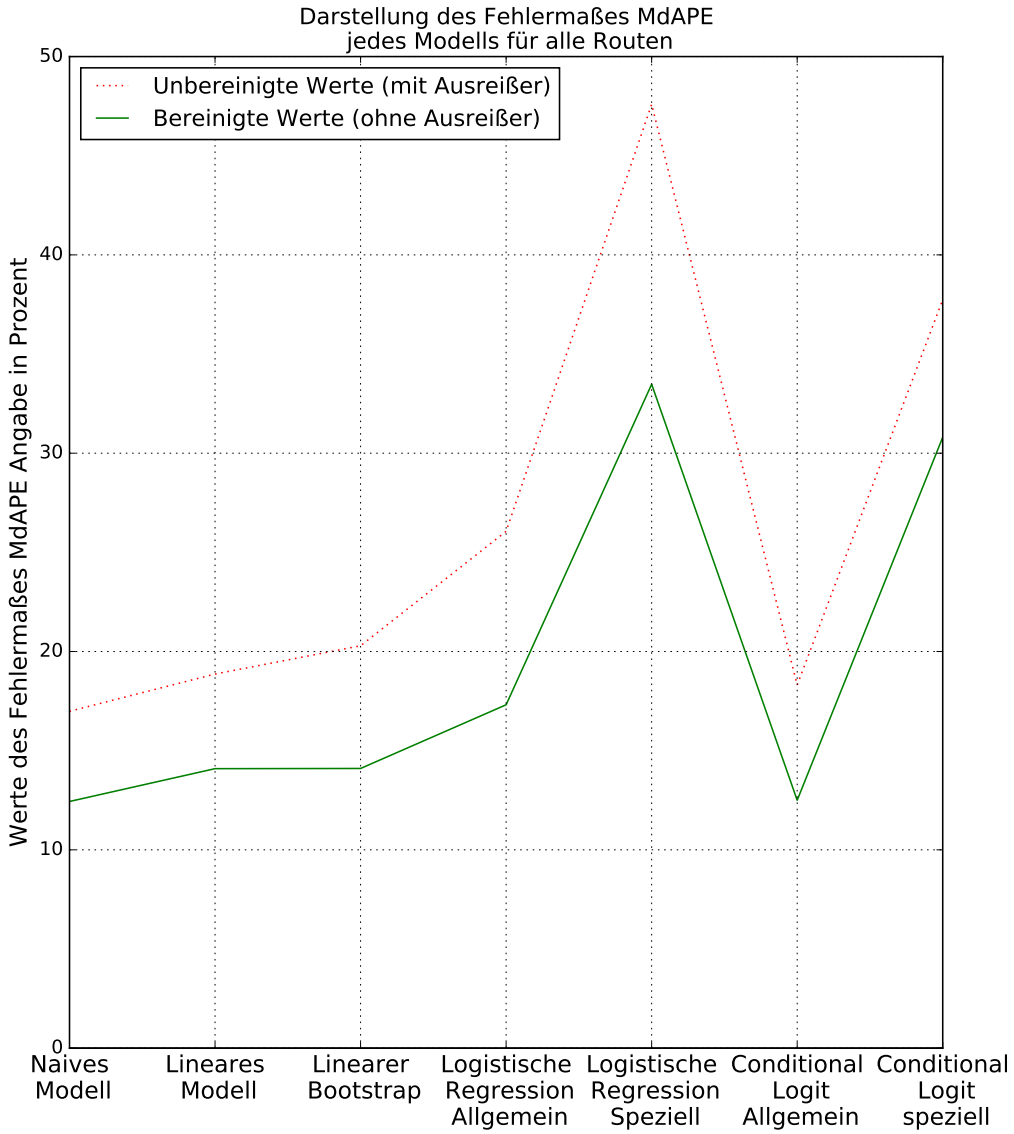


Abbildung 7.12: Darstellung des Fehlermaßes MdAPE für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

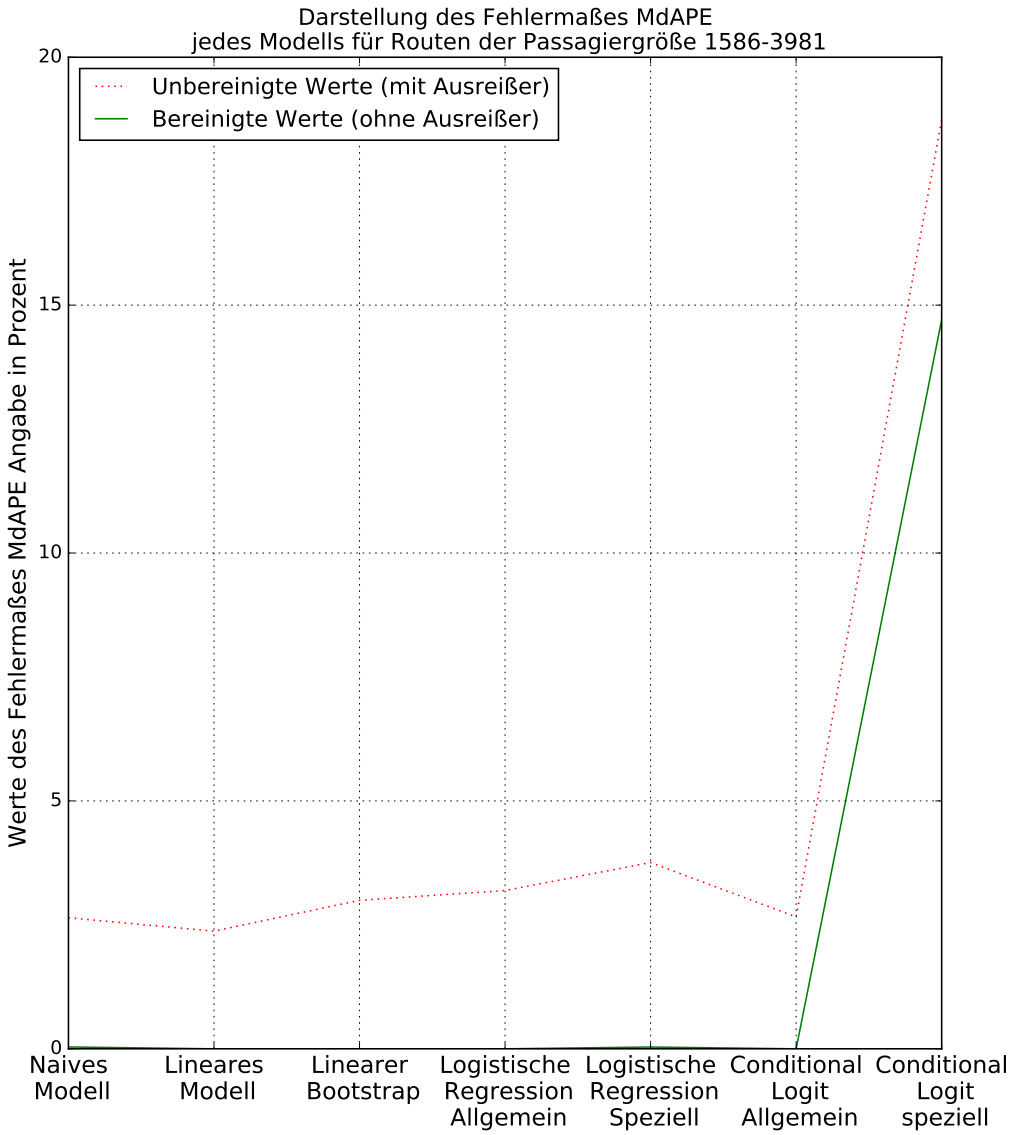


Abbildung 7.13: Darstellung des Fehlermaßes MdAPE jedes Modelles für die Passagiergrößenklasse 4: 1 586-3 981 Passagiere. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

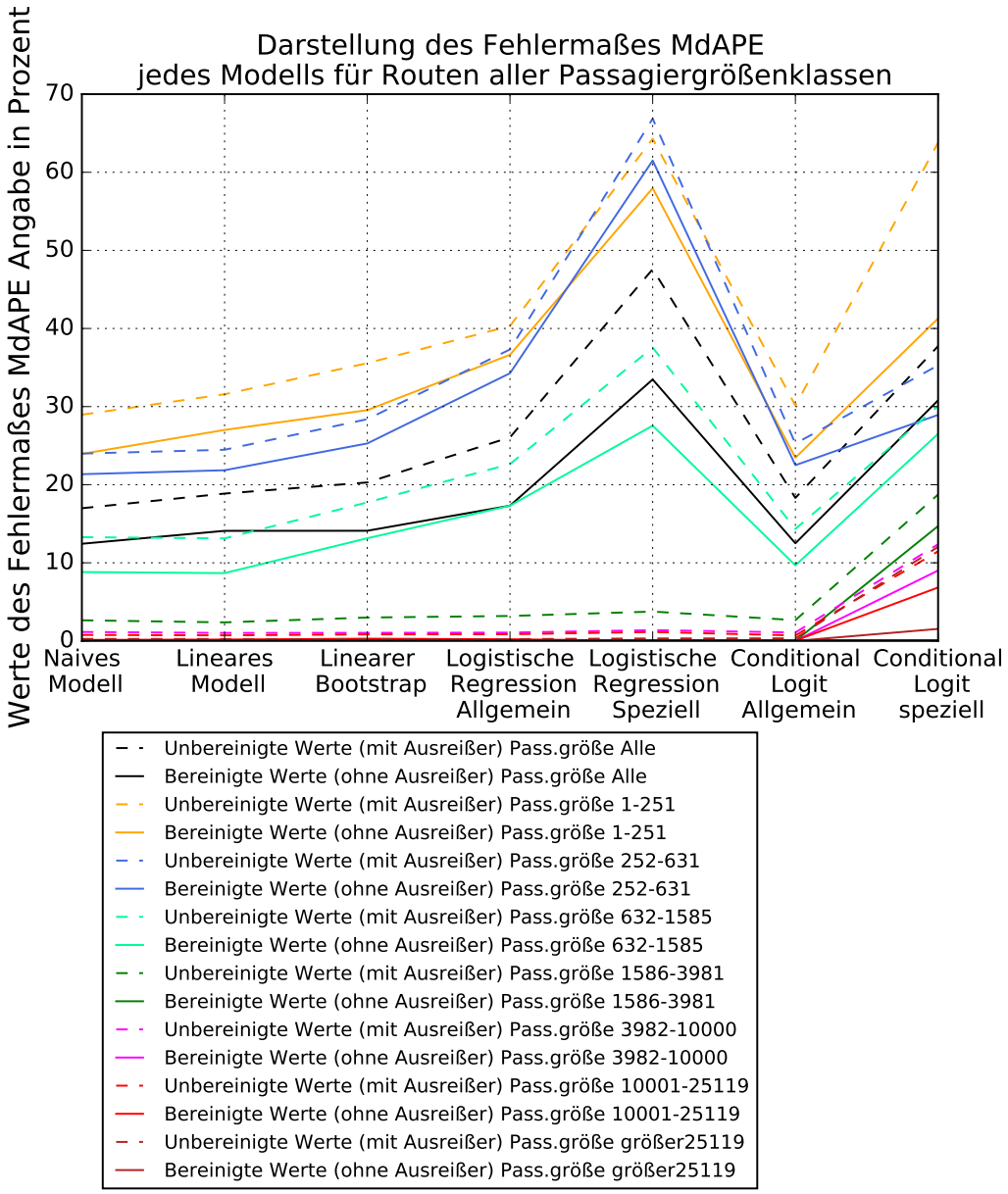


Abbildung 7.14: Darstellung des Fehlermaßes MdAPE für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.3.7 MSE - Mittlerer quadratischer Fehler

Der mittlere quadratische Fehler ist in Abbildung 7.15 und Tabelle 7.11 angegeben. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.16.

Modell	MSE mit Ausreißer	MSE ohne Ausreißer
Naives Modell	178559.2941	7741.8072
Lineares Modell	165628.1894	7607.7821
Linearer Bootstrap	305589.6919	11878.1033
Allgemeine Logistische Regression	1136208.0512	13352.1223
Spezielle Logistische Regression	2866669.7839	33404.6131
Allgemeiner Conditional Logit	8796826.9722	8468.9803
Spezieller Conditional Logit	8924042.7812	50198.6546

Tabelle 7.11: Tabelle des MSE für die Hauptmodelle, mit und ohne Ausreißern

Im Allgemeinen gleicht die Abbildung 7.15 den Abbildungen der Klassen. Hierbei ist der Conditional Logit im Vergleich zu den anderen Modellen sehr stark variant. Die logistische Regression dagegen ist nur stark variant. Die Aussage des Boxplots bestätigt sich dadurch in der Aussage, dass es in diesen Modellen große Ausreißer gibt, denn die bereinigte Version ist wesentlich weniger stark variant.
Neue Erkenntnisse liefert ein Blick in die Klassenaufteilung. Hierbei ist festzustellen, das die hohe Varianz des allgemeinen Conditional Logits bis zur Klasse 5 auftritt und danach nicht mehr. Wohingegen die spezielle Version stets eine hohe Varianz besitzt, aber erst ab Klasse 6 sowohl mit den unbereinigten als auch den bereinigten Daten sehr hohe Werte erzielt. Selbiges gilt für die logistische Regression in kleinerem Maßstab.

Das naive Modell, das lineare Modell und der lineare Bootstrap liefern im Vergleich durchgängig gute Werte.

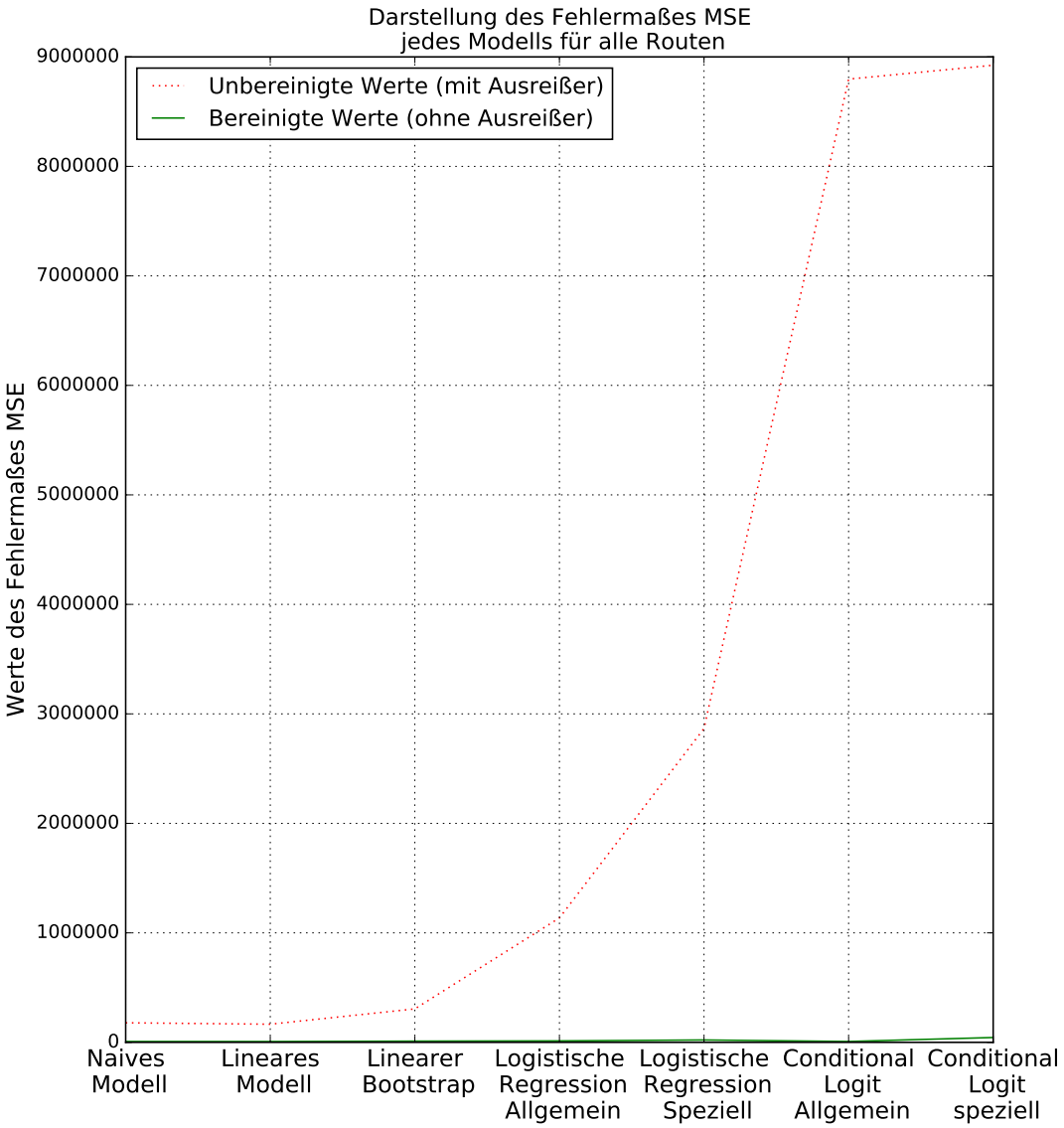


Abbildung 7.15: Darstellung des Fehlermaßes MSE für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

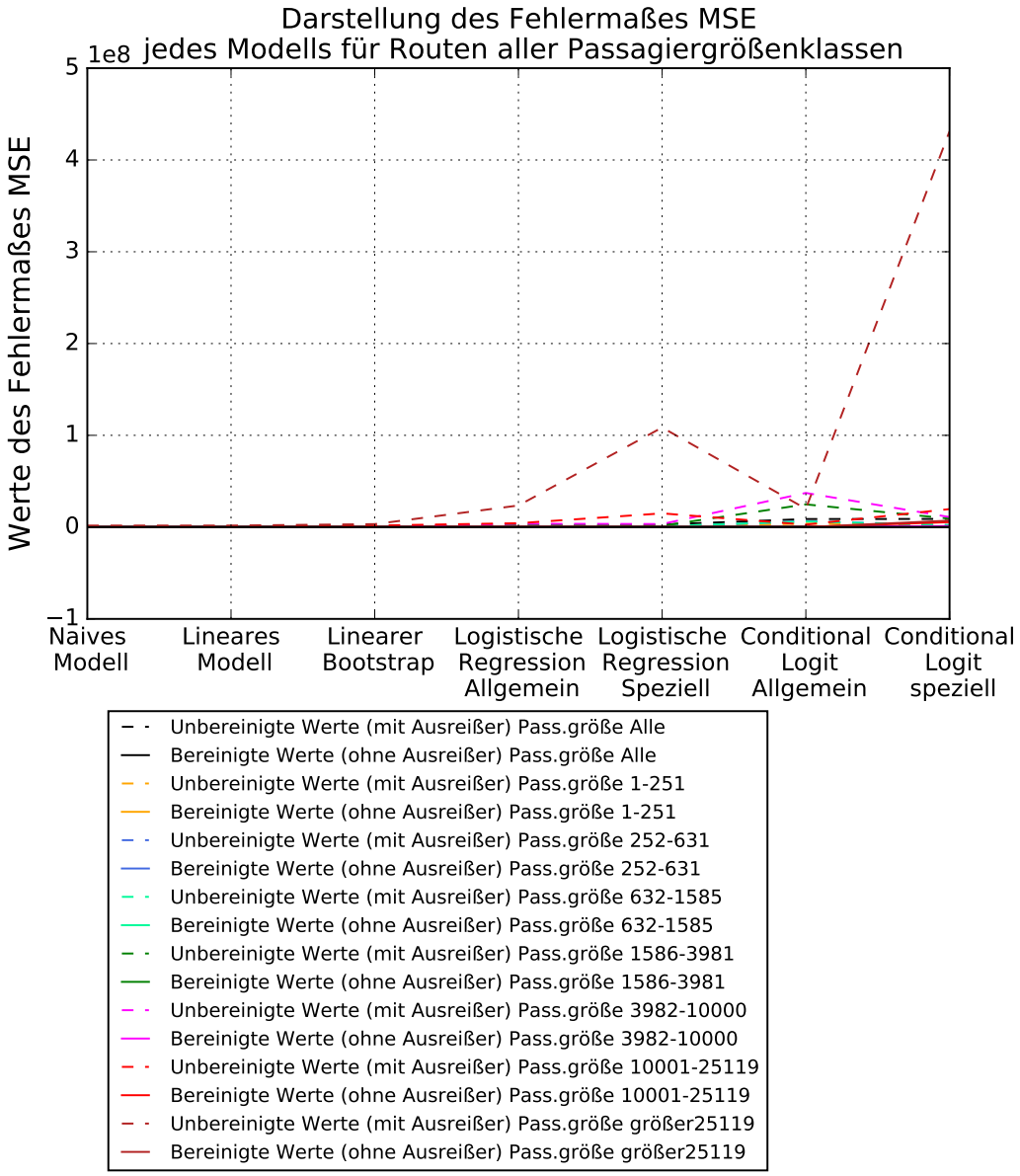


Abbildung 7.16: Darstellung des Fehlermaßes MSE für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.3.8 RMSPE - Wurzel des mittleren quadratischen prozentualen Fehlers

Der mittlere quadratische Fehler ist in Abbildung 7.17 und Tabelle 7.12 angegeben. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.18.

Modell	RMSPE mit Ausreißer	RMSPE ohne Ausreißer
Naives Modell	1.0905	0.4129
Lineares Modell	1.0907	0.4368
Linearer Bootstrap	1.3608	0.4998
Allgemeine Logistische Regression	1.8228	0.5250
Spezielle Logistische Regression	3.0543	0.6044
Allgemeiner Conditional Logit	13.0426	0.4531
Spezieller Conditional Logit	7.2556	1.1223

Tabelle 7.12: Tabelle des RMSPE für die Hauptmodelle, mit und ohne Ausreißern

Die Werte und Aussagen sind mit denen von Abschnitt 7.3.7 identisch. Ausnahme bildet hierbei der lineare Bootstrap.

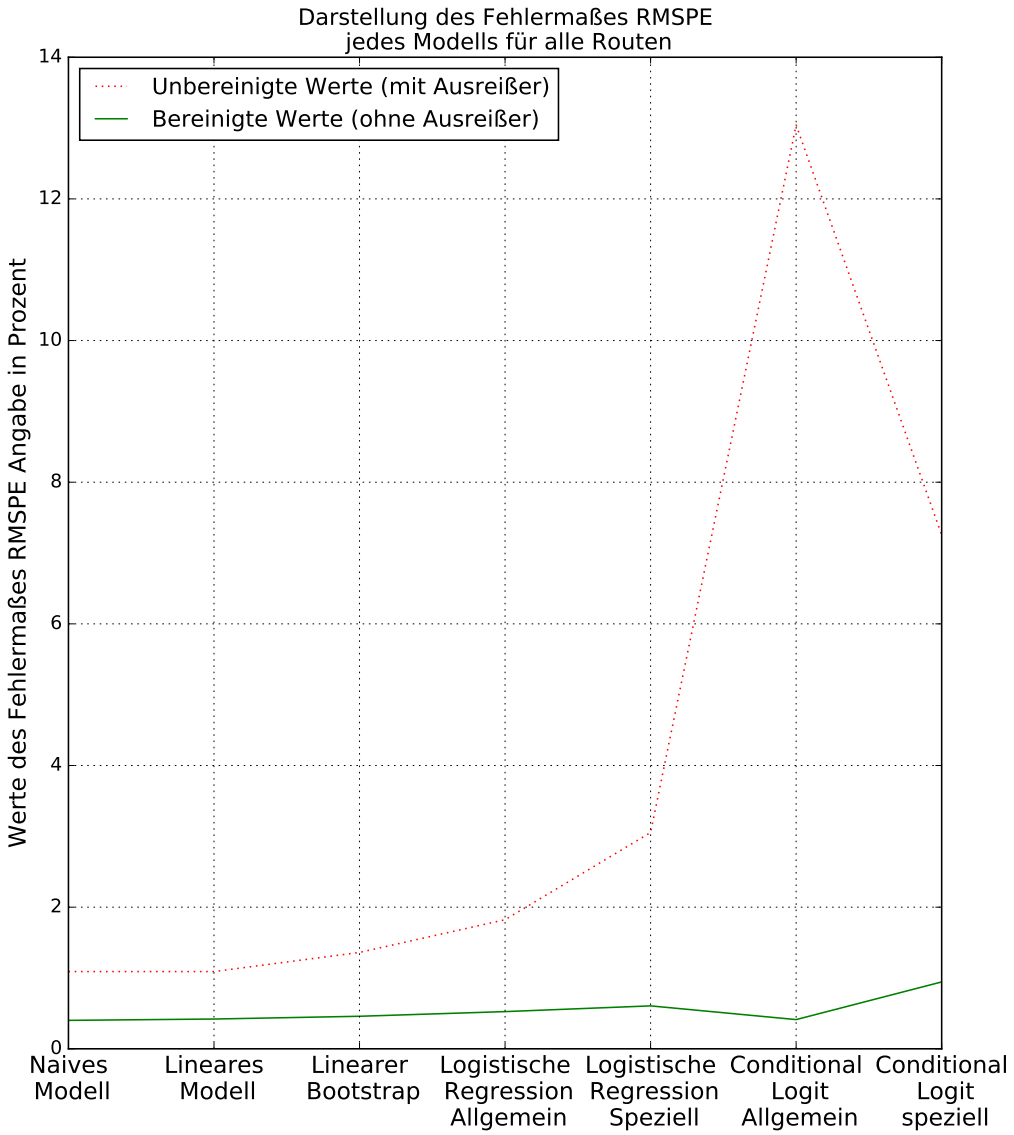


Abbildung 7.17: Darstellung des Fehlermaßes RMSPE für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

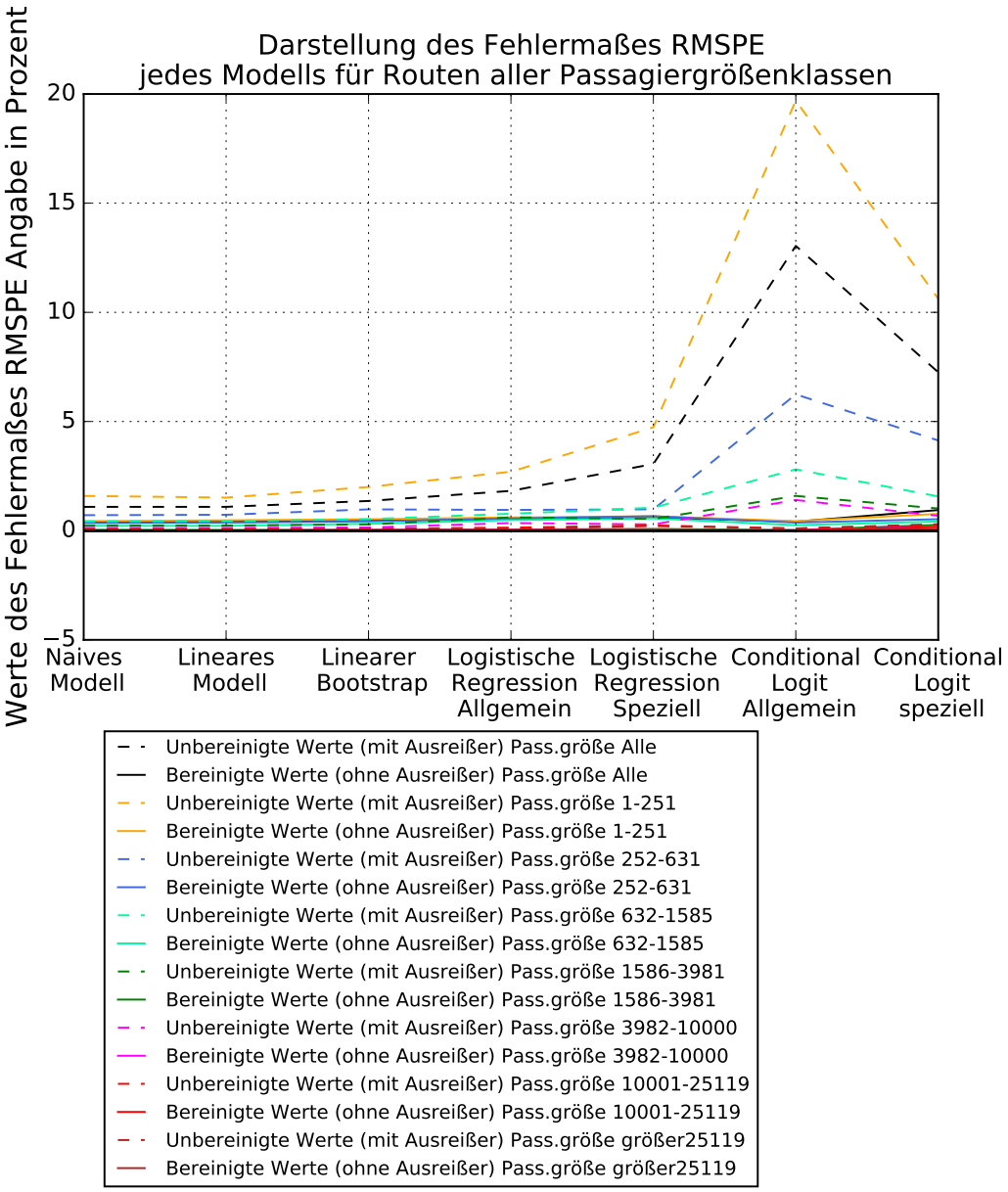


Abbildung 7.18: Darstellung des Fehlermaßes RMSPE für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.3.9 BIASMSE - Bias-Anteil des MSE

Der Bias-Anteil des mittleren quadratischen Fehlers ist in Abbildung 7.19 und Tabelle 7.13 angegeben. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.20.

Modell	BIASMSE mit Ausreißer Angabe in Prozent	BIASMSE ohne Ausreißer Angabe in Prozent
Naives Modell	0.1194	0.0076
Lineares Modell	0.3096	0.0078
Linearer Bootstrap	1.7855	0.1028
Allgemeine Logistische Regression	0.2963	0.0330
Spezielle Logistische Regression	1.5777	10.4517
Allgemeiner Conditional Logit	0.1761	0.2385
Spezieller Conditional Logit	0.00823	2.6898

Tabelle 7.13: Tabelle des BIASMSE für die Hauptmodelle, mit und ohne Ausreißern, Angabe in Prozent

Besonderes Augenmerk sollte bei diesen Werten auf die bereinigte Version gerichtet werden. Hierbei zeigen die Werte das wirkliche, vom Modell beeinflusste Verhalten der geschätzten Outputs. Während die Modelle laut den unbereinigten Daten fast gar keinen Einfluss auf den Fehler MSE besitzen ist für die bereinigte Variante teilweise genau das Gegenteil der Fall. Vor allem die spezielle logistische Regression in den Klassen 1-3 und der lineare Bootstrap, die allgemeine logistische Regression sowie der spezielle Conditional Logit zeigen in den Klassen 4-7 mit 20%-40% einen hohen Anteil des Bias am MSE. Insgesamt wirken sich die speziellen Varianten aber am stärksten aus. Diese Modelle sind dementsprechend ungeeignet für das Problem.

Der Bias steht für den Modellfehler. Der BIASMSE für den Anteil des Modellfehlers am MSE. Dementsprechend lässt sich sagen, dass die bisherigen Favoriten, das naive Modell, das lineare Modell und der allgemeine Conditional Logit vonseiten der Modellwahl als gut anzusehen sind.
Negativ überrascht hat der Bootstrap, welcher bisher eigentlich als Alternative zu den Favoriten gesehen werden konnte. Die Mittelung der Ergebnisse vielerlei Wiederholungen eines aufgeteilten linearen Modells scheint das Problem offenbar nicht ordentlich wiederzugeben. Allerdings ist anzumerken, dass der MSE des linearen Bootstrap sehr gering und der Modellfehler entsprechend als harmlos zu werten ist. Insgesamt verkleinert sich der MSE bei der bereinigten Version auch, weshalb größere BIASMSE-Werte nicht wirklich verwunderlich sind. Werden doch durch die Entfernung der Ausreißer diejenigen Werte begünstigt, welche ein Modell gut vorhersagen konnte. Somit hat ein Modell mit seinen Annahmen in der bereinigten Variante mehr Einfluss auf die erzeugten

Werte, und damit auch auf die Abweichungen.

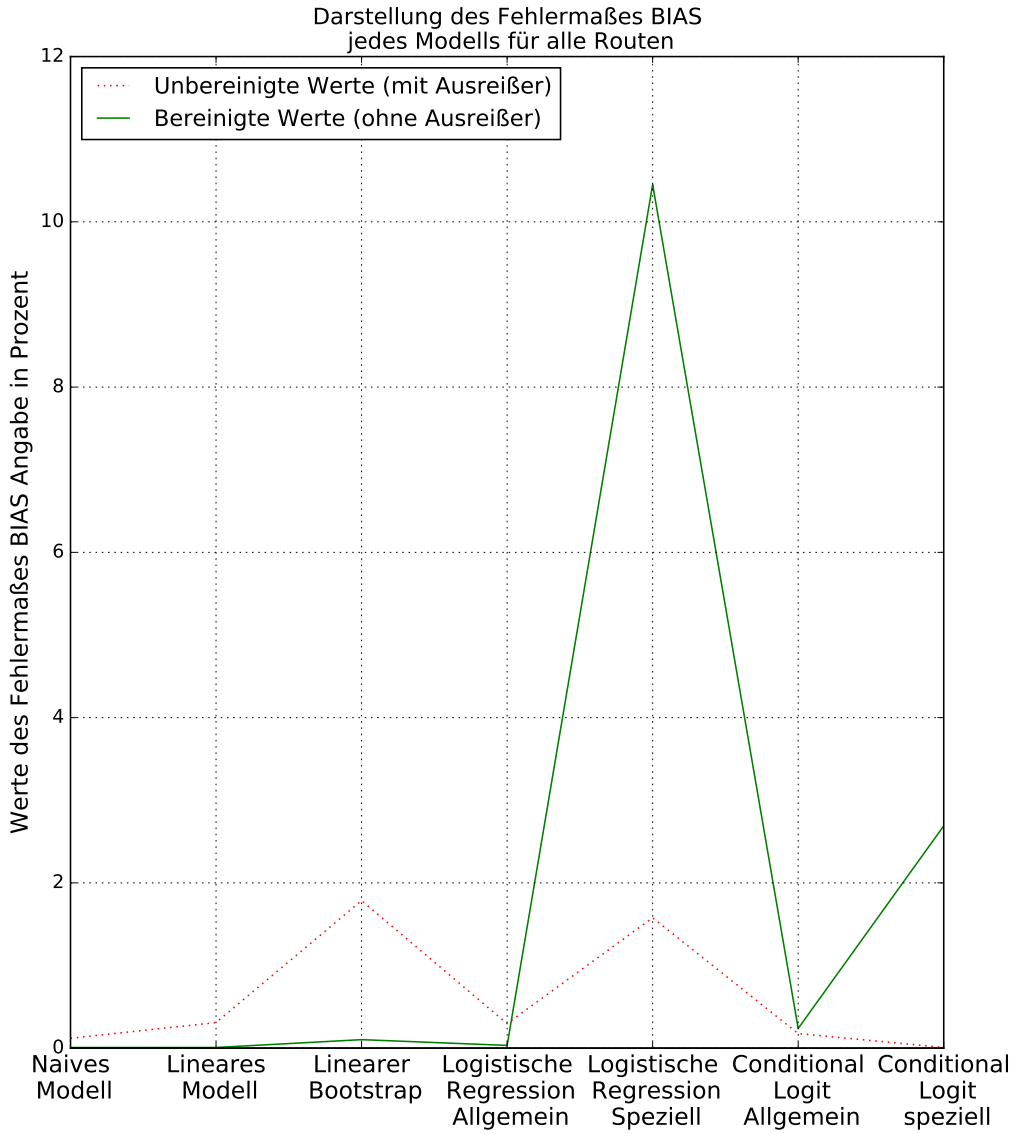


Abbildung 7.19: Darstellung des Fehlermaßes BIASMSE für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

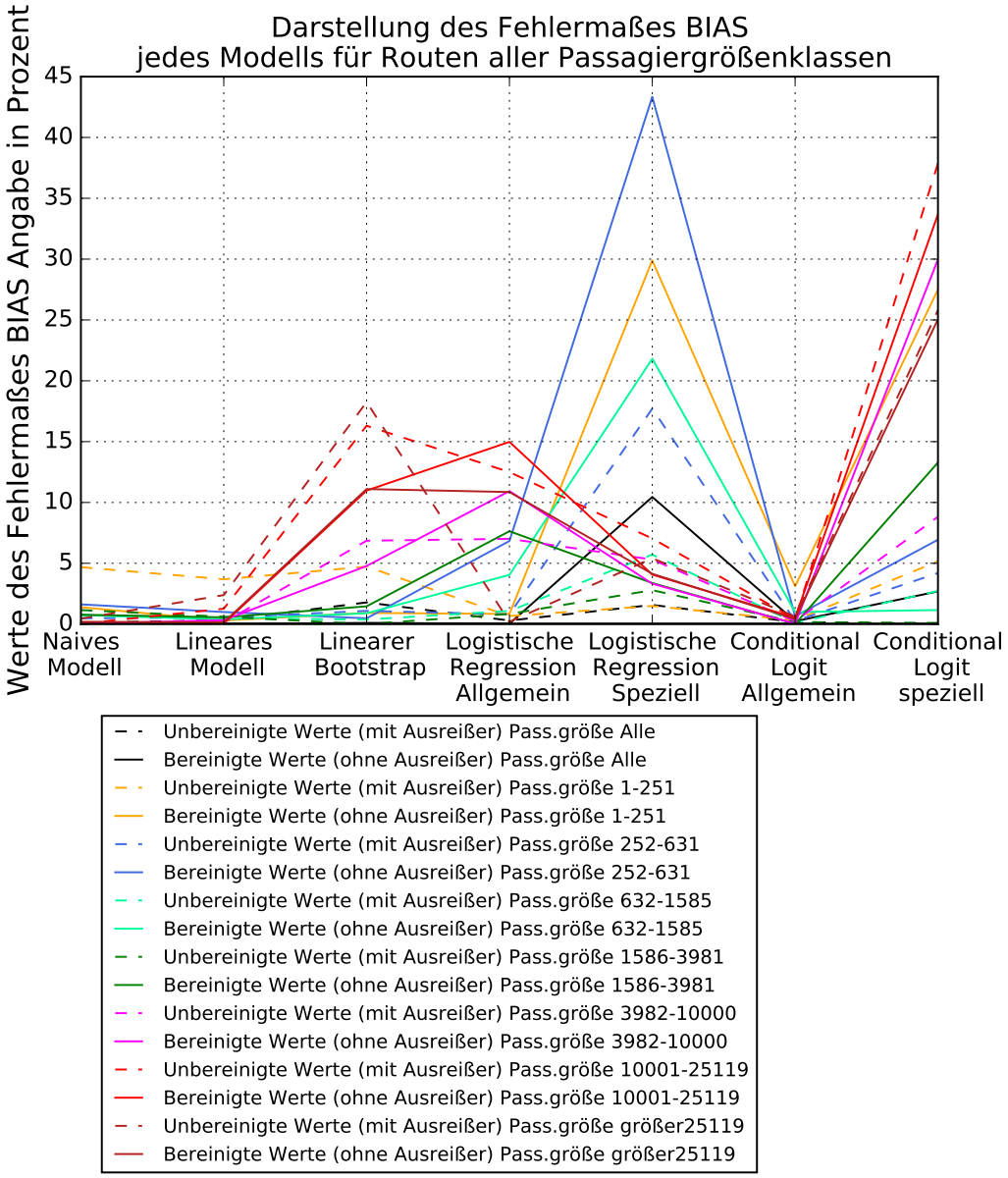


Abbildung 7.20: Darstellung des Fehlermaßes BIASMSE für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.3.10 VARMSE - Varianz-Anteil des MSE

Der Varianz-Anteil des mittleren quadratischen Fehlers ist in Abbildung 7.21 und Tabelle 7.14 angegeben. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.22.

Modell	VARMSE mit Ausreißer Angabe in Prozent	VARMSE ohne Ausreißer Angabe in Prozent
Naives Modell	0.1023	0.0084
Lineares Modell	0.5355	0.0181
Linearer Bootstrap	4.9403	0.0766
Allgemeine Logistische Regression	0.1684	0.4336
Spezielle Logistische Regression	0.6776	1.0591
Allgemeiner Conditional Logit	5.2485	0.0160
Spezieller Conditional Logit	19.5613	0.1095

Tabelle 7.14: Tabelle des VARMSE für die Hauptmodelle, mit und ohne Ausreißern

Die Aussage zu diesem Wert muss über die Klassenbetrachtung erfolgen. Allgemein bewegen sich die bereinigten Werte aller Modelle in den Klassen 1-3 im Bereich von 15%-35%. In den höheren Klassen übersteigt der VARMSE kaum 4%. Bei den unbereinigten Daten verhält sich diese Sachlage anders. Erreichen die Modelle in Klasse 6 und 7 kaum 10%, so überwinden sie in Klasse 3-5 leicht 40% und darunter sogar 60%. Dabei zeigen das naive und das lineare Modell bereits in den unteren Klassen Tendenzen zu niedrigen und damit guten Werten.

Der Varianz-Anteil des MSE steht für die Abhängigkeit der Vorhersagegenauigkeit von Änderungen der Inputdaten anteilig berechnet am MSE. Die unbereinigten Werte zeigen, dass die Ausreißer, vor allem beim Conditional Logit, einen starken Einfluss auf den MSE besitzen und seine hohen Werte zum Teil fast vollständig aus ihnen hervorgehen. Wohingegen die bereinigten Werte zeigen, dass sie den MSE fast nicht mehr beeinflussen. Die allgemeine Tendenz ist aber, dass Schwankungen in vielbereisten Routen kaum Einfluss auf die Gesamtprognosefähigkeit der Modelle besitzen. Allerdings können Routen mit wenigen Passagieren stärker streuen.

Insgesamt lässt sich aber sagen, dass die Modelle an sich wesentlich mehr Einfluss auf hohe MSE Werte haben als die Änderungen der Inputdaten. Für das lineare Modell als typischer Vertreter für Modelle mit hohem Bias und niedriger Varianz überrascht dieses Ergebnis nicht. Umso erstaunlicher ist es allerdings für die logistische Regression und den Conditional Logit, vor allem für ihre speziellen Varianten.

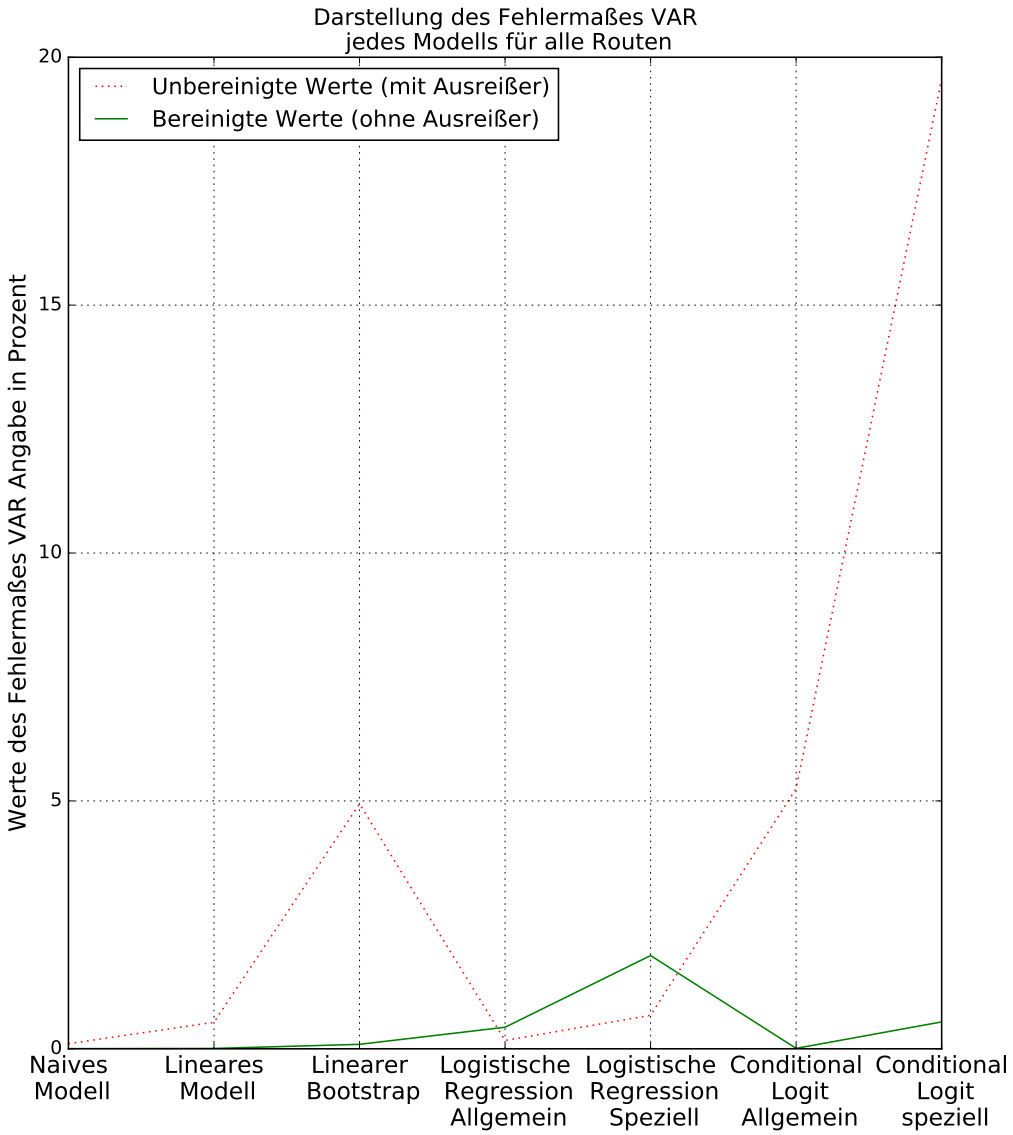


Abbildung 7.21: Darstellung des Fehlermaßes VARMSE für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

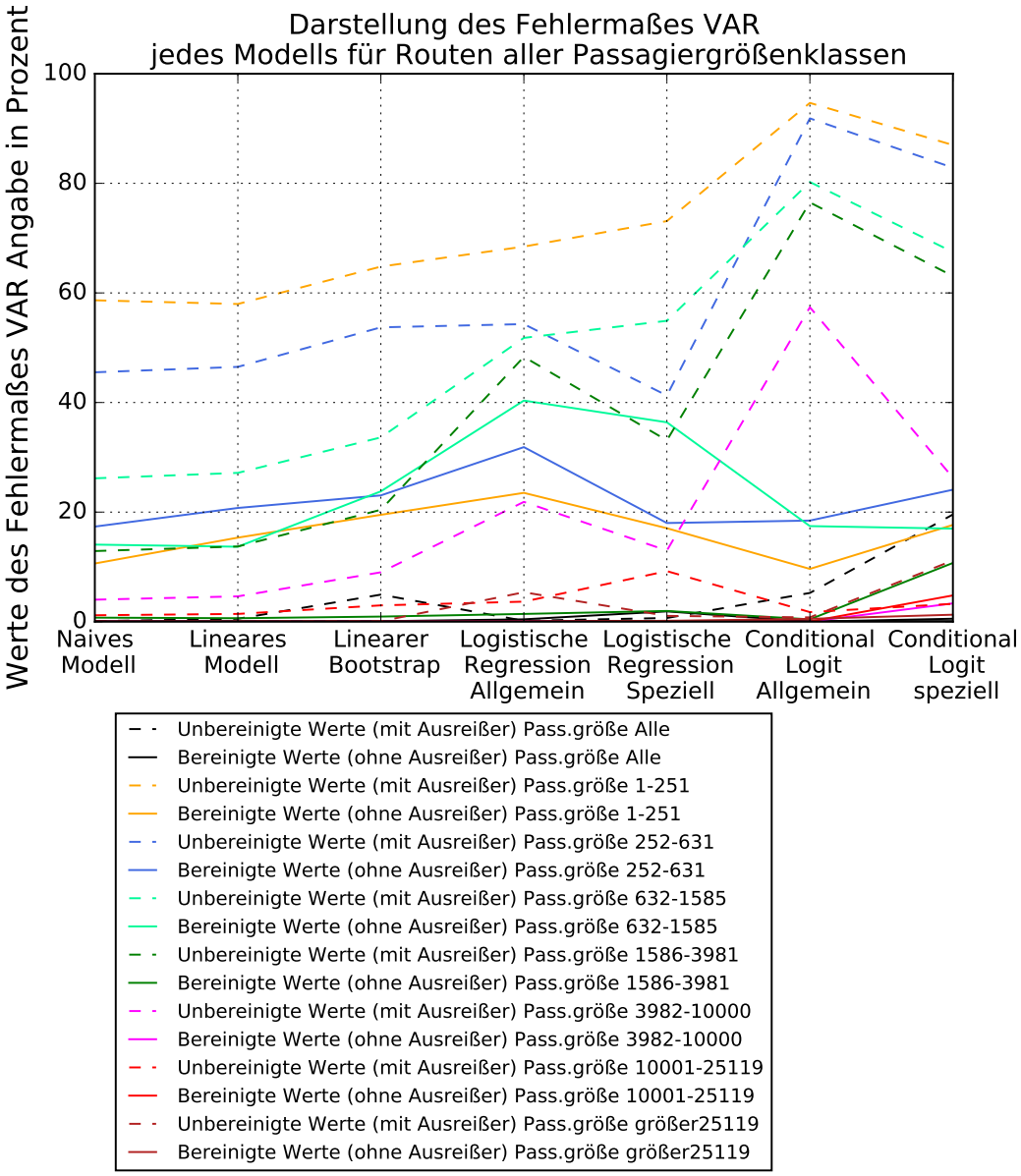


Abbildung 7.22: Darstellung des Fehlermaßes VARMSE für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.3.11 R^2 - Determinationskoeffizient

Der Determinationskoeffizient ist in Abbildung 7.23 und Tabelle 7.15 angegeben. Aufgrund des falschen Eindrucks, den die Abbildung und die Tabelle liefern, wird zusätzlich Abbildung 7.24 aufgeführt. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.25.

Modell	R^2 mit Ausreißer	R^2 ohne Ausreißer
Naives Modell	0.99734	0.99985
Lineares Modell	0.99728	0.99984
Linearer Bootstrap	0.99614	0.99979
Allgemeine Logistische Regression	0.98340	0.99974
Spezielle Logistische Regression	0.96015	0.99952
Allgemeiner Conditional Logit	0.78659	0.99964
Spezieller Conditional Logit	0.84692	0.99749

Tabelle 7.15: Tabelle des R^2 für die Hauptmodelle, mit und ohne Ausreißern

Abbildung 7.23 und Tabelle 7.15 vermitteln den Eindruck, dass für jegliche Modelle zwischen wahrer beobachteter und geschätzter Passagierzahl ein nahezu perfekter linearer Zusammenhang herrscht, welcher als nahezu perfekte Prognose gewertet werden kann. Das dies nicht so ist, zeigt Abbildung 7.24. Hierbei ist eindeutig zu sehen, dass vor allem in den unteren Klassen 1-3 eine ganz und gar nicht perfekte Beziehung herrscht. Dies erklärt sich aus der bereits ermittelten Feststellung, dass die Modelle Probleme haben, kleinere Passagierzahlen vorherzusagen.

Große Routen werden meist unterschätzt, durch ihre hohe Passagierzahl ist dies prozentual gesehen vernachlässigbar, daher ergeben sich in den hohen Klassen hervorragende Werte. Große Auswirkungen hat jedoch die tendenzielle Überschätzung kleinerer Routen. Die geschätzte Passagierzahl übersteigt oft deutlich in verschieden hohen prozentualen Anteilen die wahre beobachtete Passagierzahl (vergleiche dazu Abschnitt 7.3.5). Damit ist nicht mehr von einem linearem Zusammenhang zu sprechen. Allerdings wirkt sich die Streuung der Daten vor allem bei der logistischen Regression und dem Conditional Logit aus. Mit den bereinigten Daten ist ein guter Wert zwar schon ab Klasse 4 gegeben, durch die Ausreißer gilt dies für die allgemeinen Varianten aber erst ab Klasse 6 und die speziellen Varianten erst ab Klasse 7.

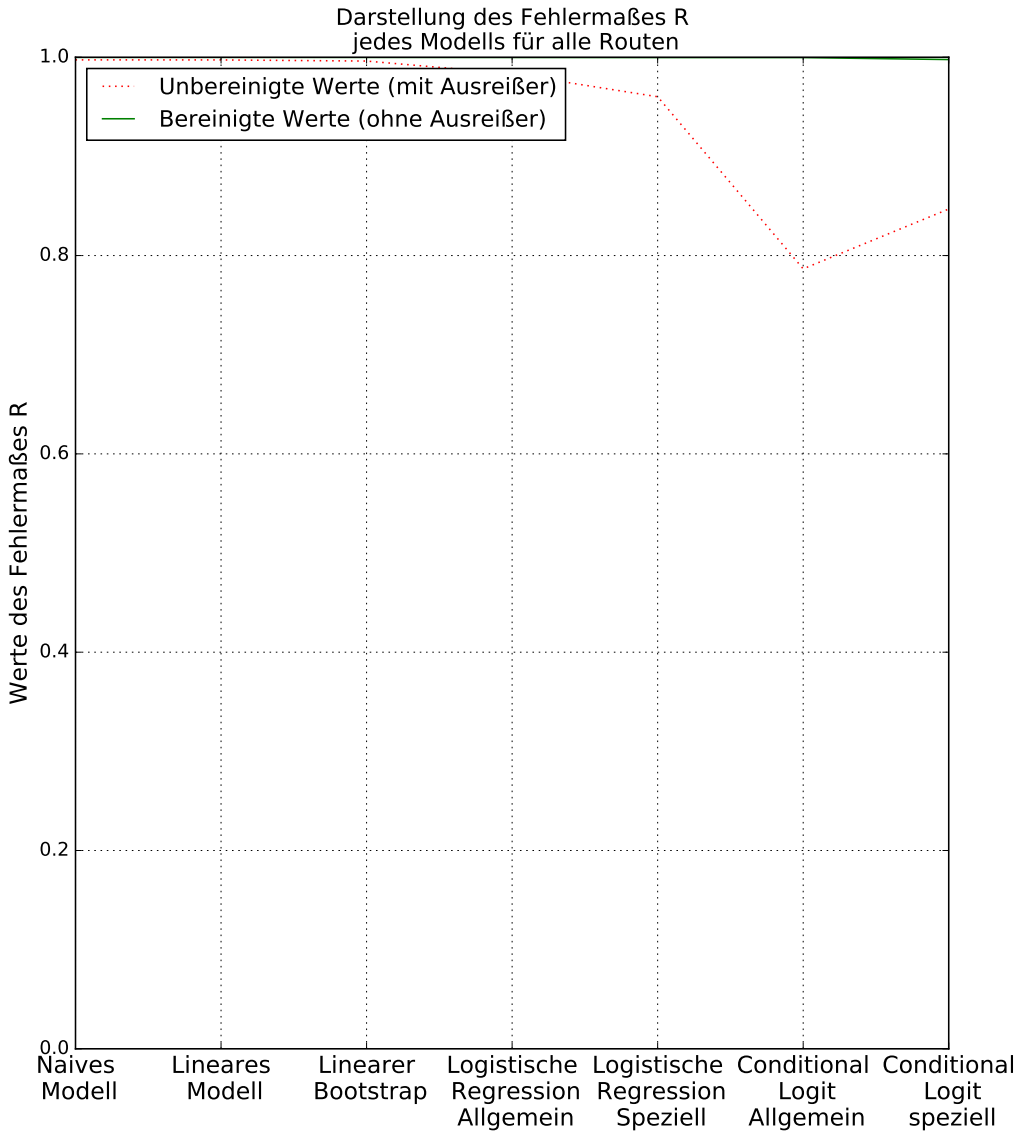


Abbildung 7.23: Darstellung des Fehlermaßes R^2 für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

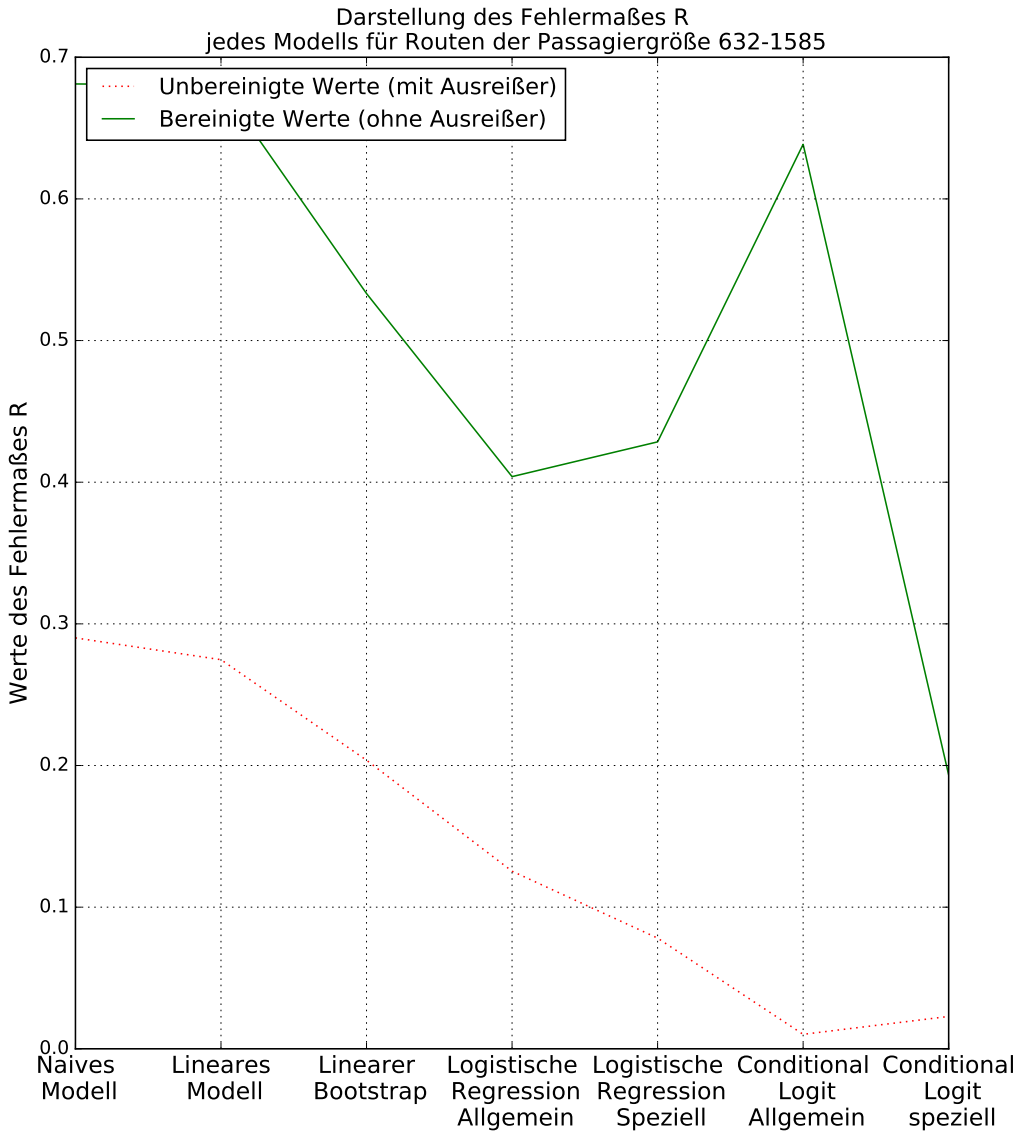


Abbildung 7.24: Darstellung des Fehlermaßes R^2 jedes Modelles für die Passagiergrößenklasse 3: 632-1 585 Passagiere. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

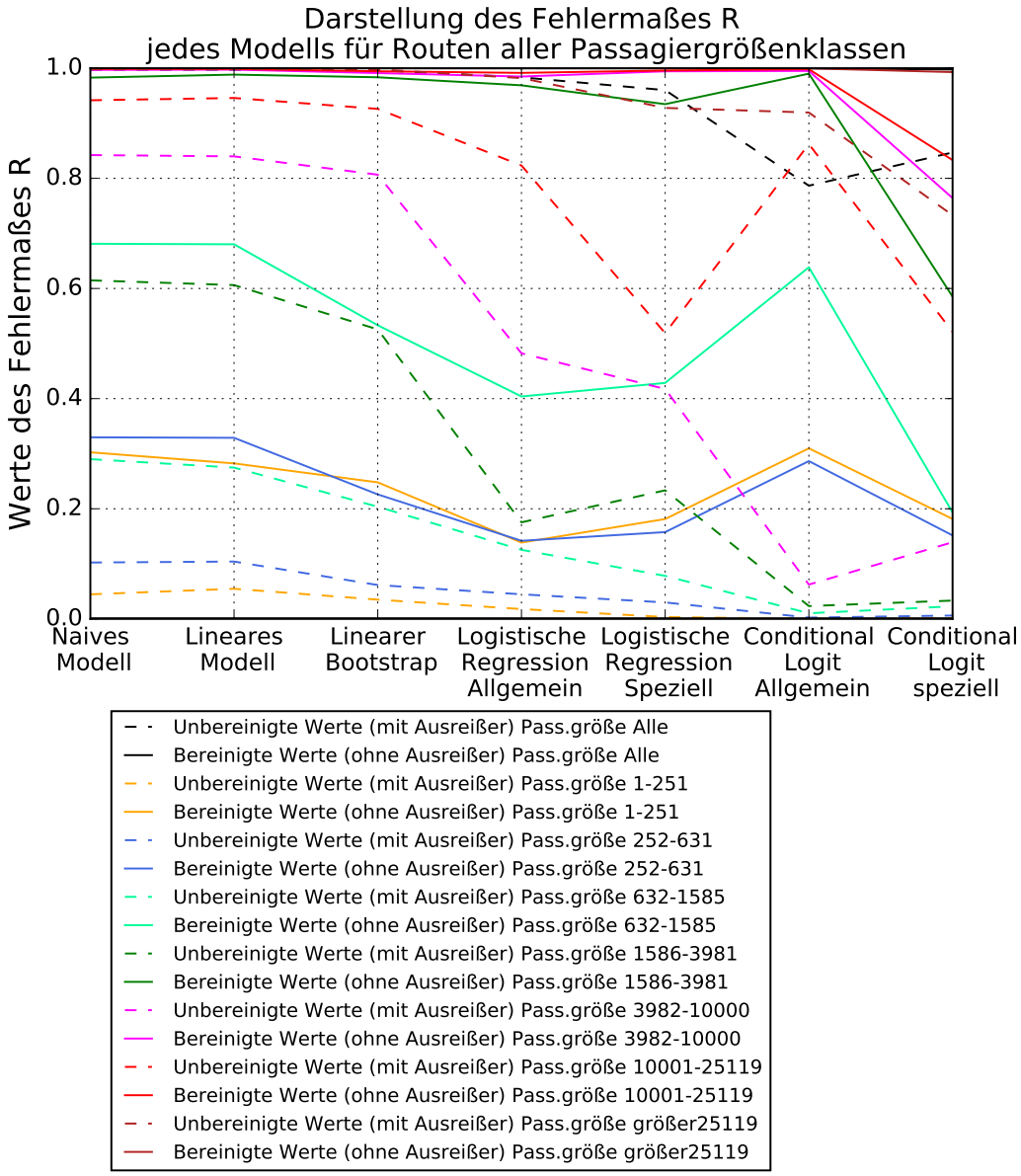


Abbildung 7.25: Darstellung des Fehlermaßes R^2 für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.3.12 KOVMSE - Kovarianz-Anteil des MSE

Der Kovarianz-Anteil des MSE ist in Abbildung 7.26 und Tabelle 7.16 angegeben. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.27.

Modell	KOVMSE mit Ausreißer Angabe in Prozent	KOVMSE ohne Ausreißer Angabe in Prozent
Naives Modell	-99.7774	-99.7593
Lineares Modell	-99.1552	-99.5804
Linearer Bootstrap	-93.2737	-97.5329
Allgemeine Logistische Regression	-99.5354	-99.5289
Spezielle Logistische Regression	-97.7456	-73.9890
Allgemeiner Conditional Logit	-94.5755	-97.1736
Spezieller Conditional Logit	-80.4301	-90.9944

Tabelle 7.16: Tabelle des KOVMSE für die Hauptmodelle, mit und ohne Ausreißern

Die Werte des KOVMSE liefern ähnliche Aussagen wie in Abschnitt 7.3.11. Hierbei ist der allgemeine Conditional Logit allerdings deutlich besser als die spezielle Variante oder die logistische Regression.

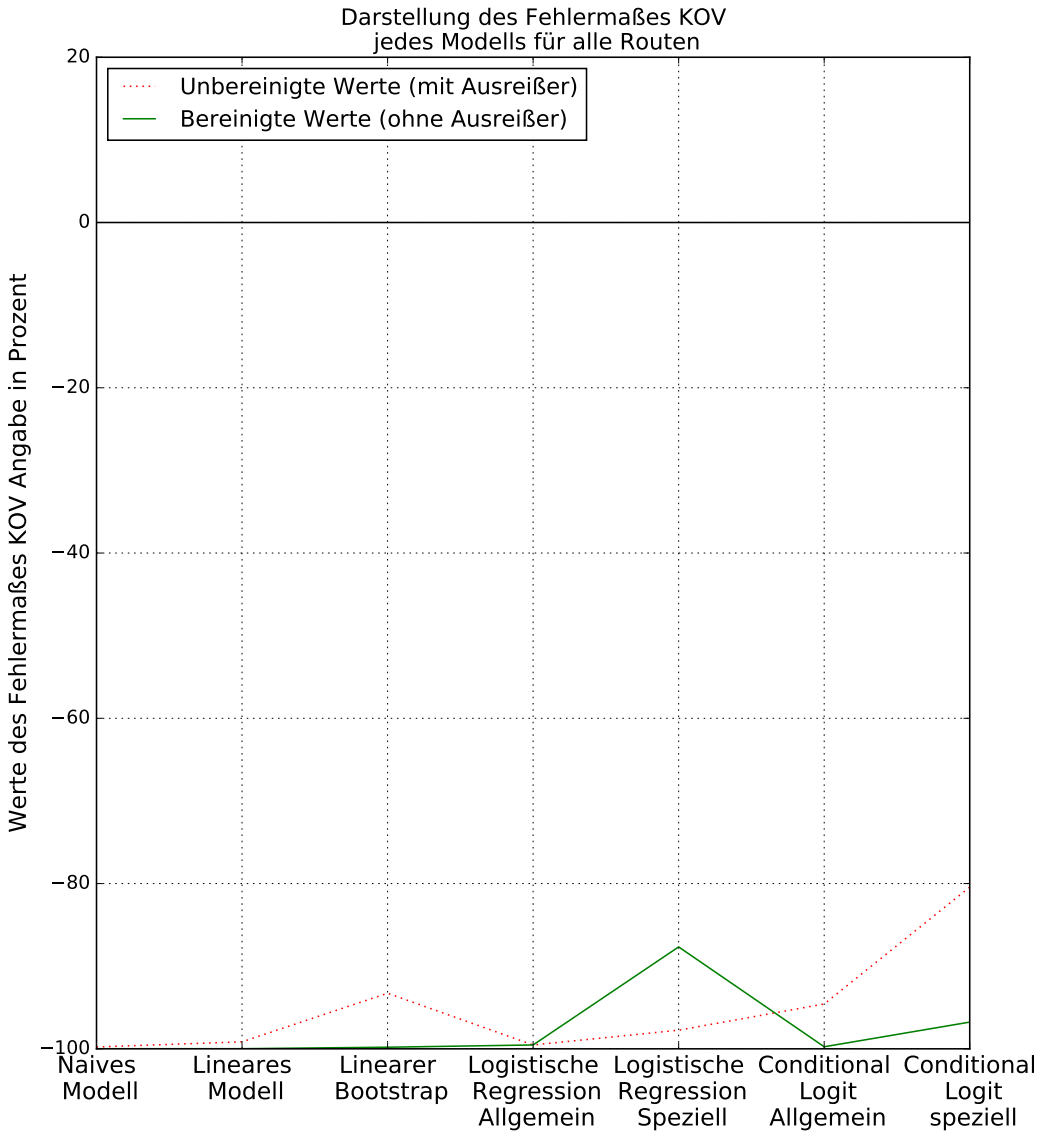


Abbildung 7.26: Darstellung des Fehlermaßes KOVMSE für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

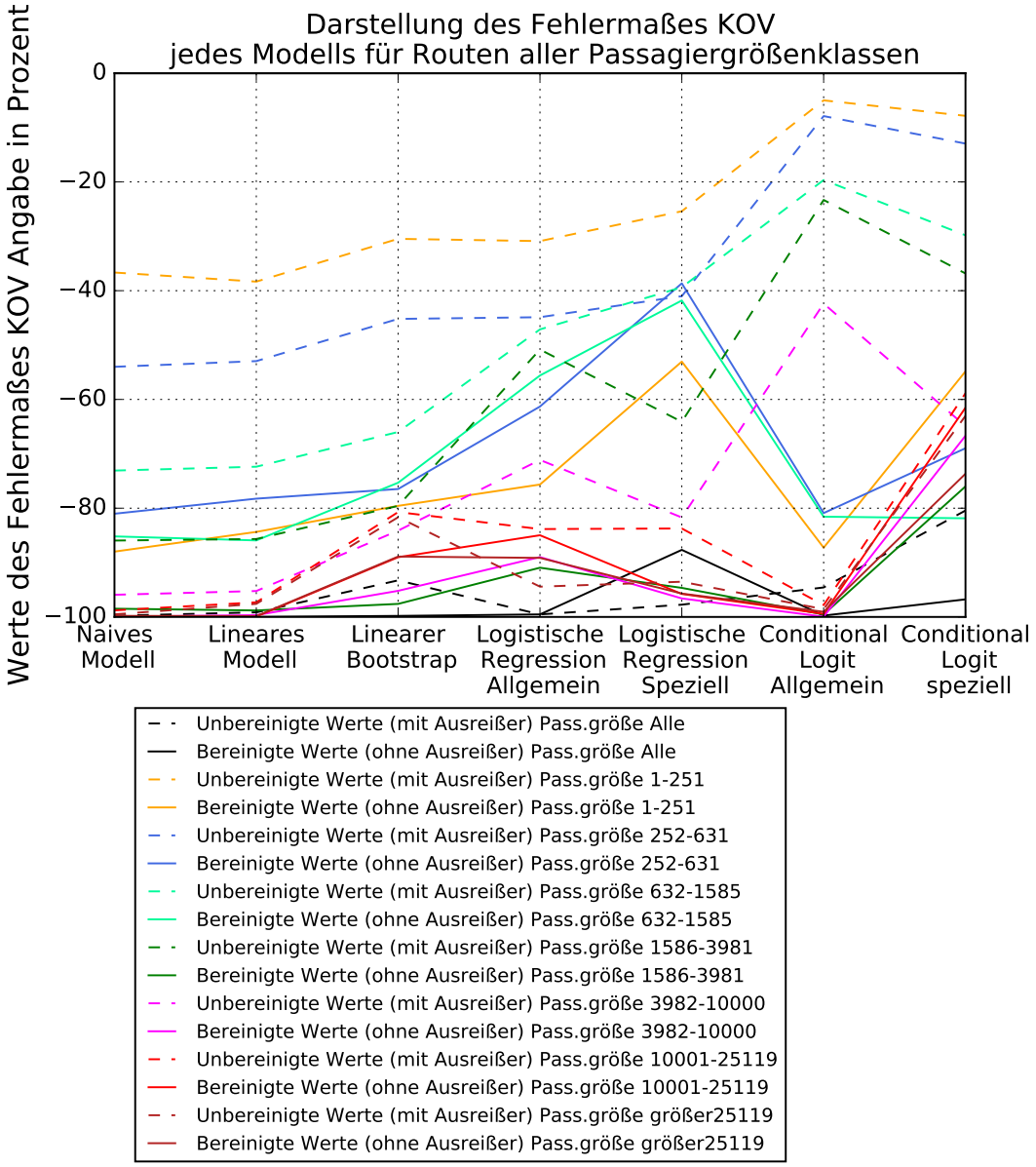


Abbildung 7.27: Darstellung des Fehlermaßes KOVMSE für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.3.13 BPC - Bravais-Pearsonsche Korrelationskoeffizient

Der Bravais-Pearsonsche Korrelationskoeffizient ist in Abbildung 7.28 und Tabelle 7.17 angegeben. Eine Abbildung der Werte aller Passagiergrößenklassen liefert Abbildung 7.29.

Modell	BPC mit Ausreißer	BPC ohne Ausreißer
Naives Modell	0.99867	0.99992
Lineares Modell	0.99864	0.99992
Linearer Bootstrap	0.99807	0.99989
Allgemeine Logistische Regression	0.99166	0.99987
Spezielle Logistische Regression	0.97987	0.99976
Allgemeiner Conditional Logit	0.88690	0.99982
Spezieller Conditional Logit	0.92028	0.99874

Tabelle 7.17: Tabelle des BPC für die Hauptmodelle, mit und ohne Ausreißern

Der BPC führt zu denselben Aussagen wie R^2 und KOVMSE, wobei der allgemeine Conditiol Logit in den bereinigten Werten stets besser ist als der lineare Bootstrap.

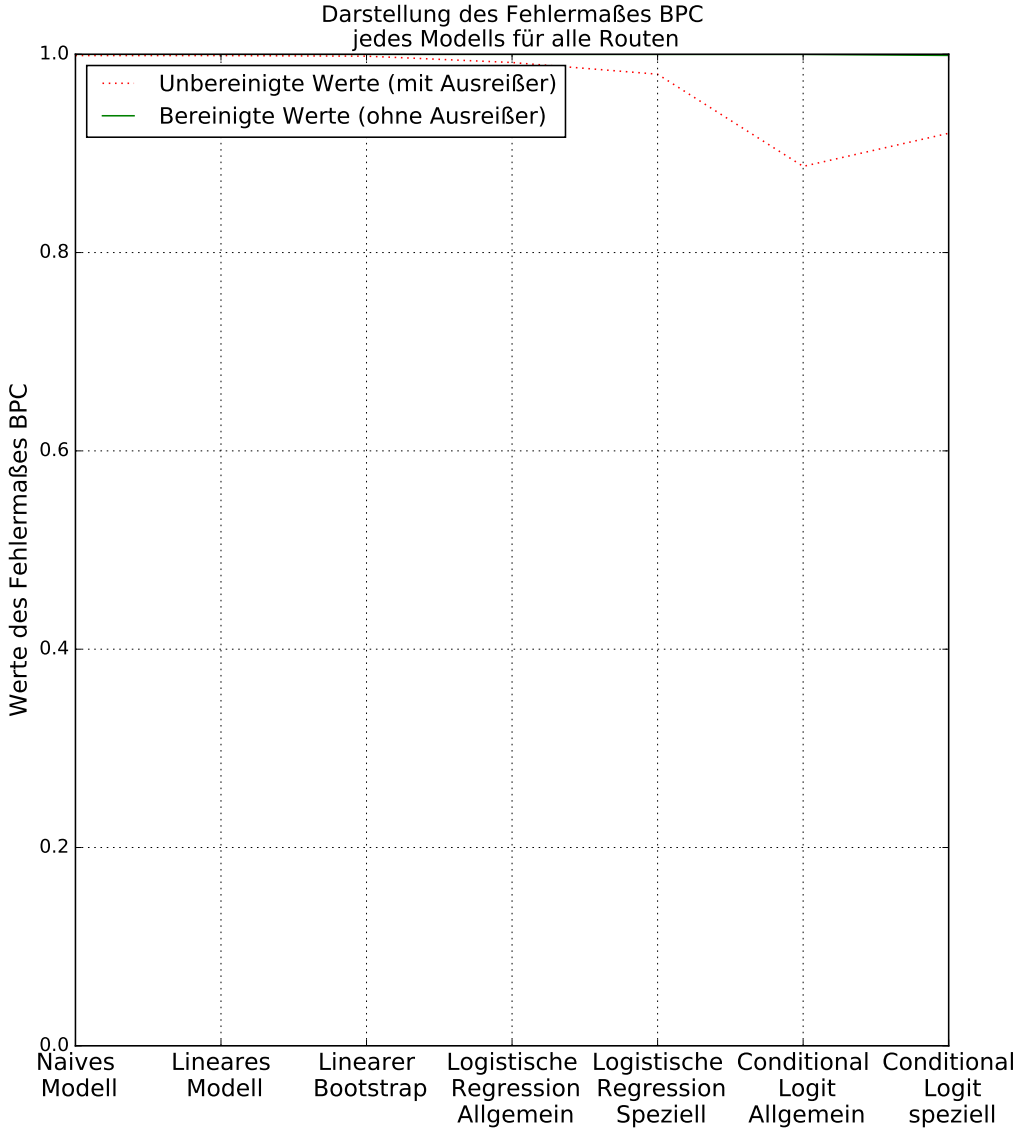


Abbildung 7.28: Darstellung des Fehlermaßes BPC für jedes Modell. Die rote Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die grüne Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

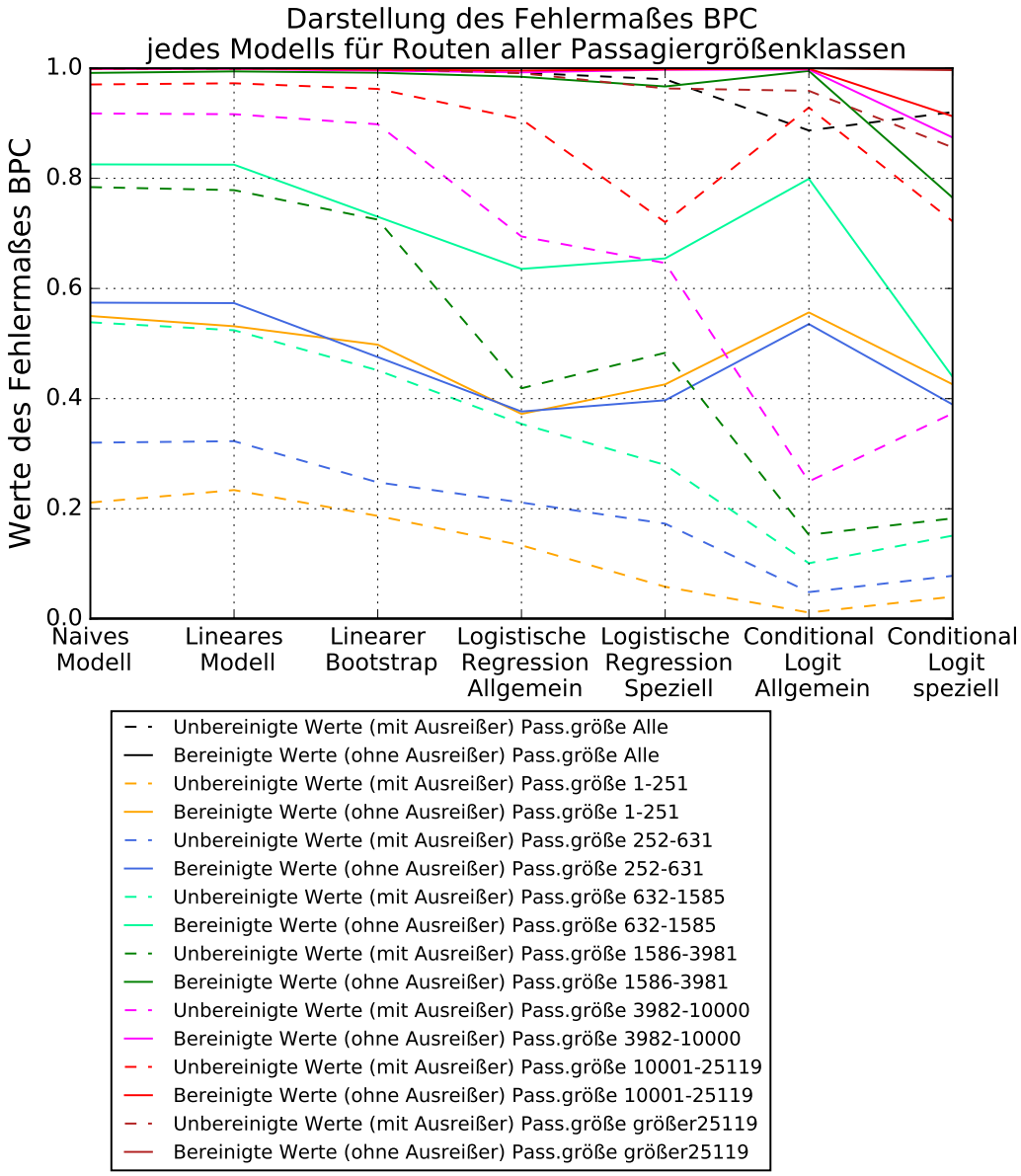


Abbildung 7.29: Darstellung des Fehlermaßes BPC für jedes Modell. In diesem Bild sind die Werte aller Passagiergrößenklassen enthalten. Für jede Klasse existieren zwei gleichfarbige Linien. Die gepunktete Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer enthalten sind. Die durchgezogene Linie gibt den Wert des Fehlermaßes an, bei dessen Berechnung die Modellausreißer nicht enthalten sind.

7.4 Fazit zu den Hauptmodellen

Abschließend lässt sich über die Verwendung der Hauptmodelle für den Einsatz der Vorhersage von Routenpassagierzahlen sagen, dass es nach den ermittelten Fehlerwerten ein eindeutiges Ergebnis gibt. Aus dem Pool der vorgestellten Möglichkeiten ist das lineare Modell die beste Wahl für die Lösung des gestellten Primärproblems. Es lieferte neben dem naiven Modell bei allen Fehlerwerten stets die besten Resultate. Es ist schnell, einfach zu verstehen und zu interpretieren. Es benötigt kaum Rechenaufwand und liefert ein $\boldsymbol{\beta}$, welches im Vergleich zum naiven Modell potenziell weniger anfällig ist für Änderungen der Daten
Diese möglichen Änderungen sind es, welche das naive Modell lediglich auf Rang zwei stellen. Es existieren keinerlei Kontrollmöglichkeiten um zum Beispiel Spitzen, hervorgerufen durch globale Ereignisse, abzufangen.
Auf Rang 3 der Anwendungsempfehlung sei der allgemeine Conditional Logit angegeben. Er besitzt vergleichbar gute Werte wie das naive Modell und das lineare Modell. Allerdings streut er einige Prognosewerte sehr stark, womit er ein Negativkriterium besitzt, welches die beiden anderen Modelle nicht haben.
Als nächstes ist der lineare Bootstrap verwendbar. Seine Werte sind mit denen des Conditional Logit und des linearen Modells vergleichbar, allerdings stets ein wenig schlechter. Außerdem benötigt er für gute Ergebnisse eine Rechenzeit, die in Stunden zu messen ist. Dies macht ihn ungeeignet für die Anwendung.
Die allgemeine logistische Regression ist den Werten nach als nächstes aufzuführen. Ihre Fehlermaße sind bereits nicht mehr mit den bereits genannten Modellen vergleichbar. Hierbei bewahrheitet sich die in Abschnitt 5.4 postulierte Befürchtung, dass das Modell aufgrund seiner Ausrichtung auf binäre Outputprobleme nicht anwendungsfähig für diese Aufgabenstellung ist.
Wie in der Auswertung sehr oft zu lesen war, sind die speziellen Varianten der logistischen Regression und des Conditional Logits als schlecht einzustufen. Vor allem der spezielle Conditional Logit. Die Idee, für jede Route ein eigenes $\boldsymbol{\beta}$ zu schätzen, hat schlichtweg nicht funktioniert. Vermutlich existiert keine ausreichende Anzahl von Daten für eine ordentliche Anpassung.
Insgesamt hat sich aber gezeigt, dass ein allgemein gehaltener Schätzvektor $\boldsymbol{\beta}$ ein gute Wahl zu sein scheint.

7.5 Auswertung zu den Fehlermaßen der Vormodelle

In den folgenden Abschnitten wird eine Auswertung für jedes Fehlermaß vorgenommen. Zur Berechnung der Fehlermaße wird der Output jeder Route, beobachtet oder prognostiziert, derartig bestimmt, dass ihm eine 2 zugeteilt wird, wenn die Route im vorhergesagten Monat auftritt. Andernfalls erfolgt die Zuordnung einer 1. Das Ergebnis einer Differenz zwischen vorhergesagtem und tatsächlichem Output beträgt 0, wenn die Aussagen zur Existenz der Route übereinstimmen. Die Differenz beträgt 1, wenn ein Auftreten einer Route prognostiziert wurde, obwohl sie in Wirklichkeit nicht vorkommt. Im gegenteiligen Fall liegt der Wert bei -1. Viele Werte 0 sind also günstig.

Bilder zu den Fehlermaßen werden nicht angegeben, da für jedes Maß lediglich 3 Werte existieren. Diese sind einfach genug zu überblicken. Für die Berechnung wurden allein diejenigen Routen herangezogen, deren Routeninformationen vollständig genug waren, um in das Modell eingegeben werden zu können.

Der Determinationskoeffizient und KOVMSE werden nicht angegeben, da ersteres Maß eigentlich für lineare Regressionen geschaffen wurde und das zweite Maß das erste enthält. Zudem ist es für das Neuronale Netz nicht berechenbar, da Werte und Durchschnittswert identisch sind. Somit ergibt sich eine Teilung durch 0.

7.5.1 Balkendiagramm des Auftretens der Routen

Abbildung 7.30 zeigt ein Balkendiagramm der drei Vormodelle bezüglich ihrer Prognosen.

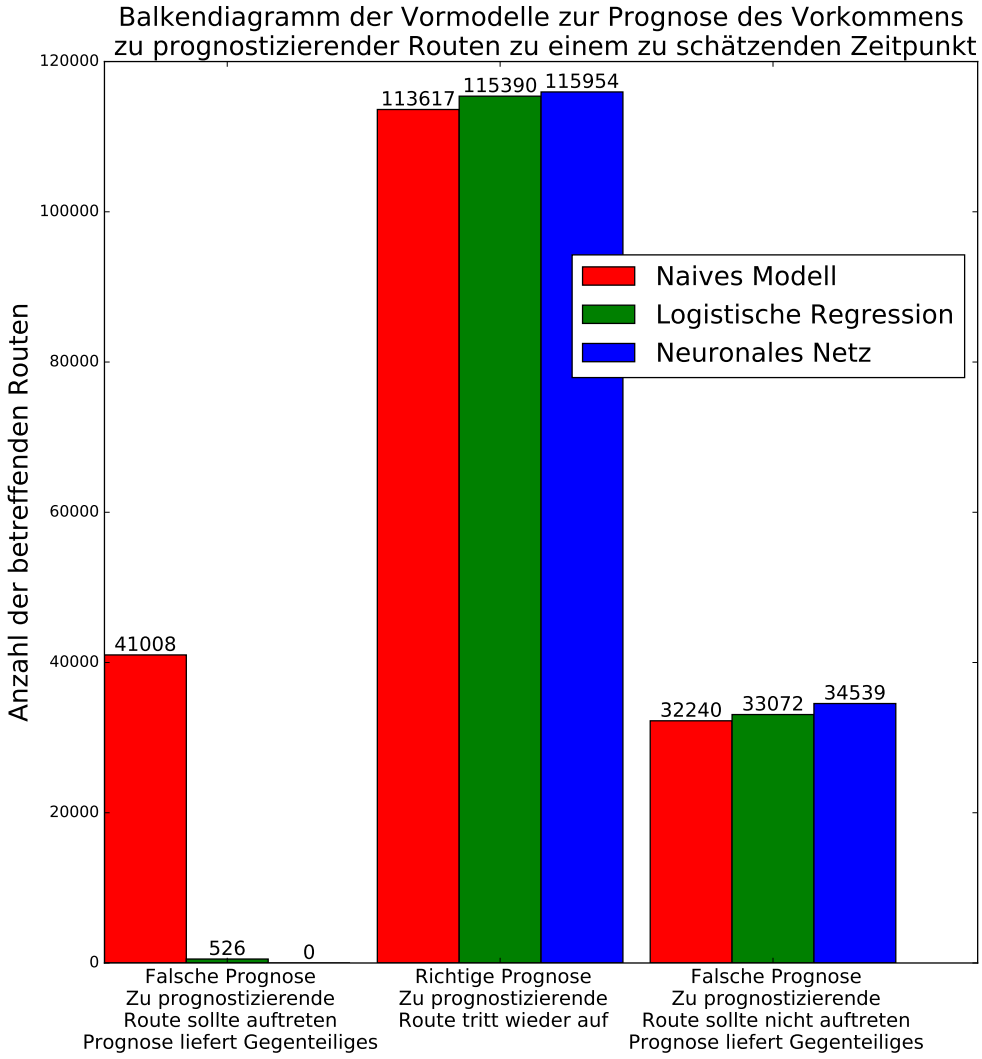


Abbildung 7.30: Darstellung der Prognosefähigkeiten des naiven Modells, der logistischen Regression und des Neuronalen Netzes. Der linke Balken zeigt die Anzahl aller Routen für welche prognostiziert wird, dass sie im zu schätzenden Monat nicht mehr auftreten, obwohl dies in Wirklichkeit doch der Fall ist. Der rechte Balken stellt ebenfalls die Anzahl aller Routen mit einer falschen Vorhersage dar. Diesmal werden alle Routen gezählt, für die das Auftreten im zu schätzenden Monat prognostiziert wird, obwohl dies in Wirklichkeit nicht der Fall ist. Der Balken in der Mitte zählt alle korrekt vorhergesagten Routen.

Um umständliche Beschreibungen zu vermeiden, soll für die nachfolgenden Abschnitte folgendes gelten:

- Die Fehlprognose, dass eine Route als nicht existent vorhergesagt wird, obwohl sie in Wirklichkeit auftritt, heißt **Negativfehler**
- Die Fehlprognose, dass eine Route als existent vorhergesagt wird, obwohl sie in Wirklichkeit nicht auftritt, heißt **Positivfehler**.

Wie zu sehen ist, liegen unter Verwendung des Referenzmonats April 2013 mehrheitlich korrekte Vorhersagen vor. Der Hauptteil der Fehlprognosen liegt bei den Positivfehlern. Anhand des roten Balkens für das naive Modell ist zu sehen, dass dies nicht unbedingt an den Modellen liegt, sondern an der Datenlage. Im Referenzmonat existieren viele Routen, welche im zu schätzenden Monat nicht auftreten. Dadurch tritt ein, was zu erwarten ist: Viele Routen werden als existent vorhergesagt, obwohl dies falsch ist.

Auffällig ist, dass die logistische Regression kaum und das Neuronale Netz gar keine Negativfehler besitzt. Dies liegt maßgeblich an der Ermittlung des Entscheidungswertes und dem Anteil der im zu schätzenden Monat auftretenden und nicht auftretenden Routen. Wie in Kapitel 6 für die jeweiligen Modelle erklärt, ist ein Entscheidungswert zu bestimmen, damit der für jede Route ermittelte Wert einer eindeutigen Aussage zugeordnet werden kann. Dabei wird für den Entscheidungswert derjenige gewählt, welcher beim Vergleich der Vergangenheitszeitpunkte den Fehler aus der Anzahl der Falschvorhersagen und der Gesamtvorhersagen minimiert. Dabei treten rund 150 000 Routen erneut auf und ca. 2 300 nicht. Daher prognostiziert das Neuronale Netz mit einem Entscheidungswert von -1, dass jede eingegebene Route wieder auftritt. Trotz des selben Vorgehens zur Bestimmung des Entscheidungswertes ist die logistische Regression in der Lage, auch auf diese minimale Menge zukünftig nicht mehr existenter Routen trainiert zu werden.

Damit erzielt das Neuronale Netz zwar keine Fehlprognosen, indem es Routen den Ausfall vorhersagt, ist aber absolut nicht flexibel und daher für die Anwendung ungeeignet. Die Entscheidung wird vollständig unabhängig von den betrachteten Wegen und Routen getroffen. Die logistische Regression besitzt hingegen die Möglichkeit zu reagieren. Erstaunlich ist, dass das naive Modell fast dieselben Werte erzielt, nur beim strittigen Punkt der Ausfallprognosen nicht. Dies lässt vermuten, dass dies diejenigen Routen sind, welche nur unzureichende Routeninformationen aufwiesen.

Es ist zu bemerken, dass die Routeninformationen der Routenparameter für die logistische Regression dieselben sind, wie für das Neuronale Netz. Die wegen mangelhafter Informationen nicht verwendbaren Routen sind also in beiden Modellen identisch. Dennoch existiert ein Unterschied in der Höhe der Positivfehler und korrekten Vorhersagen im Vergleich zu den Negativfehlern. Damit wurden mittels der logistischen Regression einige Routen korrekt als ausfallend prognostiziert. Und zwar mehr, als Negativfehler vorhergesagt wurden.

Dem Diagramm nach ist zu erwarten, dass in den Fehlermaßen gute Werte auftreten.

Diese werden womöglich eine Tendenz zeigen, Positivfehler zu erzeugen. Zudem werden die Werte für die logistische Regression und das Neuronale Netz nah beieinander liegen.

7.5.2 ME - Mittlerer Fehler

Die Werte des ME sind in Tabelle 7.18 angegeben.

Modell	ME
Naives Modell	-0.0469
Logistische Regression	0.2184
Neuronales Netz	0.2295

Tabelle 7.18: Tabelle des ME für die Vormodelle

Das naive Modell liefert gute Werte. Allerdings nur, weil sich Negativ- und Positivfehler ausgleichen. Die Tendenz neigt allerdings zu einer ausgeglichenen Bilanz. Bei nur drei Werten bedeutet dies eine recht gleichmäßige Streuung der Fehler. Demzufolge ist die logistische Regression und das Neuronale Netz schlechter als das naive Modell, bei Betrachtung des Balkendiagramms 7.30 kann von Rechtsschiefe gesprochen werden.

7.5.3 MAE - Mittlerer absoluter Fehler

Die Werte des MAE sind in Tabelle 7.19 angegeben.

Modell	MAE
Naives Modell	0.3919
Logistische Regression	0.2255
Neuronales Netz	0.2295

Tabelle 7.19: Tabelle des MAE für die Vormodelle

Bald 40 Prozent aller Vorhersagen des naiven Modells sind falsch, wohingegen die beiden anderen Modelle nur eine halb so große Fehlerrate aufweisen. Die logistische Regression ist dabei geringfügig besser als das Neuronale Netz. Dies ist allerdings zu erwarten, da das naive Modell die Fehler gleichmäßig streut und die anderen Modelle fast ausschließlich Positivfehler besitzen.

7.5.4 MPE - Mittlerer prozentualer Fehler

Die Werte des MPE sind in Tabelle 7.20 angegeben.

Modell	MPE Angabe in Prozent
Naives Modell	6.2804
Logistische Regression	22.0212
Neuronales Netz	22.9505

Tabelle 7.20: Tabelle des MPE für die Vormodelle

Der mittlere prozentuale Fehler ist dem ME nach beim naiven Modell sehr gering. Die logistische Regression und das neuronale Netz hingegen konnten ihren Wert vom MAE beinahe beibehalten. Nur die wenigen Negativfehler der logistische Regression verringern den Wert ein wenig.

7.5.5 MAPE - Mittlerer absoluter prozentualer Fehler

Die Werte des MAPE sind in Tabelle 7.21 angegeben.

Modell	MAPE Angabe in Prozent
Naives Modell	28.2257
Logistische Regression	22.3742
Neuronales Netz	22.9505

Tabelle 7.21: Tabelle des MAPE für die Vormodelle

Nach den Werten des MAE ist erstaunlich, dass die Fehlklassifikationsrate des naiven Modells derart gut zu sein scheint. Die anderen Modelle bestätigen ihre Werte.

7.5.6 MdAPE - Median des absoluten prozentualen Fehlers

Die Werte des MdAPE sind in Tabelle 7.22 angegeben.

Modell	MdAPE Angabe in Prozent
Naives Modell	0
Logistische Regression	0
Neuronales Netz	0

Tabelle 7.22: Tabelle des MdAPE für die Vormodelle

Dieses Ergebnis war abzusehen und überrascht nicht. Immerhin wird das Auftreten der meisten Routen korrekt vorhergesagt. Dies hat damit zu tun, dass eine einmal eingeführte Route kein kurzzeitiges Phänomen darstellt, sondern aufgrund einer gewissen regelmäßigen Dringlichkeit/Nachfrage nötig/günstig wurde. Ausnahmen sind hierbei meist globale Großereignisse, welche Passagierspitzen erzeugen, die mit der normalen Flugroutenkapazität nicht kompensiert werden können.

7.5.7 MSE - Mittlerer quadratischer Fehler

Die Werte des MSE sind in Tabelle 7.23 angegeben.

Modell	MSE
Naives Modell	0.3919
Logistische Regression	0.2255
Neuronales Netz	0.2295

Tabelle 7.23: Tabelle des MSE für die Vormodelle

Diese Werte entsprechen genau denen des MAE, da als Differenzenwerte lediglich -1,0,1 entstehen.

7.5.8 RMSPE - Wurzel des mittleren quadratischen prozentualen Fehlers

Die Werte des RMSPE sind in Tabelle 7.24 angegeben.

Modell	RMSPE Angabe in Prozent
Naives Modell	0.4768
Logistische Regression	0.4720
Neuronales Netz	0.4790

Tabelle 7.24: Tabelle des RMSPE für die Vormodelle

Hierbei überrascht das naive Modell erneut mit einem nahezu identischen Wert zu denen der beiden anderen Modelle. Der Trend der Annäherung der prozentualen Fehler setzt sich fort.

7.5.9 BIASMSE - Bias-Anteil des MSE

Die Werte des BIASMSE sind in Tabelle 7.25 angegeben.

Modell	BIASMSE Angabe in Prozent
Naives Modell	0.5616
Logistische Regression	21.1607
Neuronales Netz	22.9505

Tabelle 7.25: Tabelle des BIASMSE für die Vormodelle

Der durch das Modell zu verantwortende Fehler beim MSE ist beim naiven Modell bemerkenswert gering. Üblicherweise ist dies so zu interpretieren, dass das Modell die Natur der Problemstellung äußerst genau trifft. Der Fehler des MSE liegt damit nicht am Modell.
Die beiden anderen Modelle bestätigen lediglich ihre Fehlerrate.

7.5.10 VARMSE - Varianz-Anteil des MSE

Die Werte des VARMSE sind in Tabelle 7.26 angegeben.

	Angabe in Prozent
Modell	VARMSE
Naives Modell	0.3312
Logistische Regression	56.2833
Neuronales Netz	77.0494

Tabelle 7.26: Tabelle des VARMSE für die Vormodelle

Die Erklärung aus Abschnitt 7.5.9 zum naiven Modell gilt auch hier. Da Varianz und Bias bezüglich des MSE äußerst gering sind, entfallen die restlichen 99% auf das weiße Rauschen, welches durch das naive Modell nicht erklärt werden kann. Aufgrund seiner Modellannahmen ist dies nicht verwunderlich.
Das Neuronale Netz hingegen besitzt fast kein weißes Rauschen. Der hohe VARMSE Wert zeigt, dass es stark anfällig für Schwankungen in den Daten ist. Dies erklärt sich mit seiner bereits eruierten Unflexibilität aufgrund seines Entscheidungswertes von -1.
Die logistische Regression ist ebenfalls recht unflexibel, wie bereits in Abschnitt 7.5.1 ausgeführt wurde. Dies ist allerdings nicht die Schuld des Modells sondern der Datenlage. Da es aber dennoch fähig ist, das Nichtauftreten von Routen vorherzusagen, ist sein VARMSE-Wert deutlich geringer als der des Neuronalen Netzes.

7.5.11 BPC - Bravais-Pearsonsche Korrelationskoeffizient

Die Werte des BPC sind in Tabelle 7.27 angegeben.

Modell	BPC
Naives Modell	-0.2421
Logistische Regression	-0.0317
Neuronales Netz	Nicht berechenbar

Tabelle 7.27: Tabelle des BPC für die Vormodelle

Der BPC zeigt an, dass die Daten der originalen und der geschätzten Routen nichtlinear zusammenhängen. Bei der logistischen Regression ist dies nicht verwunderlich. Das naive Modell überrascht mit einer leichten Negativkorrelation schon eher. Allerdings zeigt ein Wert von -0.24 lediglich einen schwachen linearen Zusammenhang an.

7.6 Fazit zu den Vormodellen

Wie bereits erwähnt, ist das Neuronale Netz aufgrund seiner Inflexibilität nicht zur Anwendung zu empfehlen. Des Weiteren ist aufgrund einiger überraschender Werte der Fehlermaße für das naive Modell zu überlegen, ob für das behandelte Klassifikationsproblem nicht geeignetere Fehlermaße existieren.
Allein von der prozentualen Fehlerrate her ist die logistische Regression dem naiven Modell vorzuziehen. Sie liegt auch in den Vorgaben des DLR. Des Weiteren ist es der logistischen Regression möglich, den äußerst geringen Anteil an Routenausfällen abzubilden. Das naive Modell kann dies nicht, oder ist in diesem Punkt stark von den gewählten Zeitpunkten abhängig. Zudem ist eine erfolgreiche Langzeitvorhersage bei der logistischen Regression wahrscheinlicher als beim naiven Modell.

Auffällig ist die Übereinstimmung des naiven Modells und der logistischen Regression in den Anzahlen der korrekten Vorhersagen und der Positivfehler. Da diese Anzahl auch mit dem unflexiblen Modell des Neuronalen Netzes übereinstimmt, scheint eine Eigenschaft für das Nichtauftreten einer Route in mangelhaften Routeninformationen zu liegen. Vermutlich sind die entsprechenden Segmente, Flugzeuge oder Flughäfen zu unbedeutend, um in irgendeiner namhaften Datenbank zu erscheinen. Auf jeden Fall ist dieses Datenverhalten auffällig und untersuchungswürdig.

Für die Anwendung wird die logistische Regression empfohlen.

8 Zusammenfassung

Die vorliegende Arbeit widmete sich der Aufgabe, eine Problemstellung des Deutschen Zentrums für Luft- und Raumfahrt mit den Mitteln der Mathematik zu lösen. Für eine bekannte Anzahl an Flugpassagieren sollte ihr Start- und Zielflughafen für einen nahen zukünftigen Zeitpunkt bekannt sein. Ein Paar aus Start- und Zielflughafen wird auch Weg genannt. Zielstellung war es, die Aufteilung dieser Passagiere auf unterschiedliche Flugrouten zu prognostizieren.
Für diese Aufgabe wurden 5 verschiedene Modelle der Statistik auf ihre Anwendbarkeit bezüglich Geschwindigkeit, Prognosegüte und Anpassung an die spezifische Problemstellung hin untersucht. Das Naive Modell, das Lineare Modell, der Lineare Bootstrap, die Logistische Regression und ihre Erweiterung, der Conditional Logit.
Eine sekundäre Problemstellung bestand darin, die Existenz der benötigten Flugrouten sowie deren Daten für den Vorhersagezeitpunkt (oder einen anderen beliebigen Zeitpunkt) zu prognostizieren. Hierbei kamen das Naive Modell, die Logistische Regression und das Modell des künstlichen Neuronalen Netzes zum Einsatz.

Nach einer Einführung der Problemstellung, der Vorstellung und Aufbereitung der Daten und der Erörterung einiger wichtiger Hintergundinformationen der Statistik begann eine ausführliche Beschreibung der Modelle. Dabei wurde festgestellt, dass sowohl für die logistische Regression, als auch für den Conditional Logit eine allgemeine und eine spezielle Variante existiert. Die allgemeine Version verwendet eine Modifikation des Modells, sodass es auf alle Wege angewendet werden kann. Die speziellen Modelle ermitteln für jeden Weg eine eigene angepasste Einstellung.

Letztendlich stellte sich heraus, dass diese speziellen Varianten nicht für die Anwendung geeignet sind. Die Abweichungen in der Prognose sind, verglichen mit den anderen Modellen, zu stark. Weiterhin wurde bereits in der theoretische Erarbeitung festgestellt, dass die logistische Regression nicht für Problemstellungen mit multiplen Lösungsmöglichkeiten geeignet ist. Daher wurde auch der Conditional Logit untersucht, welcher eine Erweiterung der logistischen Regression hinsichtlich dieses Problems ist.
In der Auswertung trat deutlich zutage, dass sowohl das naive, als auch das lineare Modell die besten Resultate hinsichtlich fast aller verwendeten Fehlermaße liefert. Diese Ergebnisse liegen sogar deutlich innerhalb des vom Deutschen Zentrum für Luft und Raumfahrt vorgegebenen Rahmens. Allerdings ist das naive Modell als einfache Übertragung vergangener Datensätze auf neue Zeitpunkte als nicht theoretisch zuverlässig einzuschätzen. Auch der Conditional Logit liefert vergleichbar hervorragende Ergebnisse. Diese werden allerdings von einigen sehr stark ausreißenden Prognosewerten verwässert. Der lineare Bootstrap ist nicht einsetzbar, da er eine sehr hohe Rechenzeit besitzt. Die beste Alternative ergibt sich also mit dem linearen Modell, dessen Implementation innerhalb weniger Minuten eine hervorragende Lösung liefert.

Bei der Erarbeitung zur Lösung des Sekundärproblems ergab es sich, dass es äußerst schwierig ist, den Ausfall einer Flugroute zu prognostizieren, da dies nur äußerst selten der Fall ist. Das neuronale Netz stellte sich selbst darauf ein, alle Flugrouten als existent zu bewerten. Dies ist nicht akzeptabel. Das naive Modell besitzt wiederum überhaupt keine Möglichkeit, auf diesen Umstand zu reagieren. Allein die logistische Regression war in der Lage, sich an diesen kleinen Anteil an Flugroutenausfällen anzupassen. Sie ist daher zur Lösung dieses Problems zu empfehlen. Das Modell liefert innerhalb weniger Minuten ein Ergebnis, welches sich knapp im Rahmen der Vorgaben befindet.

Fragen und offene Probleme existieren für die primäre und sekundäre Aufgabenstellung hinsichtlich der benötigten und verwendeten Eingabeinformationen. Es würde spezieller Analysen wie Leave-One-Out, Lasso etc. bedürfen, um festzustellen, welche Informationen über eine Flugroute relevant und welche irrelevant sind. Für das lineare Modell ist dies weniger interessant als für die logistische Regression und seine Erweiterung, welche diese Daten benötigen. Zudem könnten weitere Untersuchungen und Überlegungen hinsichtlich zusätzlicher Inputinformationen angestellt werden.
In diesem Rahmen wurde eine Überlegung interessant, welche während der Auswertung der Sekundärmodelle auftrat. So scheint die Nichtexistenz von Eingabeinformationen ebenfalls als Information fungieren zu können. Dies könnte ein interessanter Gesichtspunkt in eben erwähnten Untersuchungen und Überlegungen sein.
Während der Auswertung der Sekundärmodelle traten zudem einige seltsame Werte in den Fehlermaßen auf, welche nicht zufriedenstellend erklärt werden konnten. Es wäre zu untersuchen, ob die verwendeten Fehlermaße für dieses Klassifikationsproblem überhaupt geeignet sind und ob nicht bessere Vergleichskriterien existieren.
Weiterhin ist für das Sekundärproblem anzumerken, das die logistische Regression allein aufgrund der schlechten Eignung der anderen Modelle empfohlen wird. Ein Vergleich mit anderen Klassifikationsmodellen wäre anzustellen.
Weiterhin wurde nicht untersucht, wie sich die Ergebnisse der Lösung des sekundären Problems auf die Güte der Lösung des Primärproblems auswirken.

Zusammenfassend soll gesagt sein, dass die Lösungen der gestellten Aufgaben zufriedenstellend ausfallen.

Literaturverzeichnis

[AAS] https://airandspace.si.edu/exhibitions, letzter Aufruf: 24.05.2017

[ACA] http://www.acare4europe.com/documents/latest-acare-documents/acare-flightpath-2050, letzter Aufruf: 24.05.2017

[AGR02] A. Agresti, Categorical Data Analysis, Second Edition, JOHN WILEY & SONS, 2002

[AGR07] A. Agresti, An Introduction to Categorical Data Analysis, Second Edition, JOHN WILEY & SONS, 2007

[ANA] A. D. Anastasiadis, G. D. Magoulas, M. N. Vrahatis, New globally convergent training scheme based on the resilient propagation algorithm, Neurocomputing 64, Seite 253-270, 2005

[AIR] http://www.airliners.de/amadeus-marketingdaten-emirates-finnair-taca/20178, letzter Aufruf: 22.05.2017

[APB] http://airportsbase.org, letzter Aufruf: 22.05.2017

[ATH] http://www.stat-athens.aueb.gr/ jpan/diatrives/Motakis/chapter7.pdf, letzter Aufruf: 23.05.2017

[ARE10] T. Arens, F. Hettlich, C. Karpfinger, U. Kockelkorn, K. Lichtenegger, H.Stachel, Mathematik, Springer Verlag, 2010

[BAH] D. Bahdra, Choice of Aircraft Fleets in the US NAS: Findings from a Multinomial Logit Analysis, Center for Advanced System Development (CAASD), https://www.mitre.org/sites/default/files/pdf/bhadra_analysis.pdf, letzter Aufruf: 23.05.2017

[BAH03-1] D. Bahdra, J. Gentry, B. Hogan, M. Wells, CAASD's Future Air Traffic Timetable Estimator: A Micro-Econometric Approach, The MITRE Corporation's Center for Advanced Aviation System Development (CAASD), 2003, https://www.mitre.org/sites/default/files/pdf/bhadra_estimator.pdf, letzer Aufruf: 23.05.2017

[BAH03-2] D. Bhadra, DEMAND FOR AIR TRAVEL IN THE UNITED STATES: BOTTOM-UP ECONOMETRIC ESTIMATION AND IMPLICATIONS FOR FORE-

CASTS BY ORIGIN AND DESTINATION PAIRS, Center for Advanced Aviation System Development (CAASD), Journal of Air Transportation Vol. 8, No. 2, 2003

[BAS] https://wwz.unibas.ch/fileadmin/wwz/redaktion/fai/MSc_Oekonometrie_FS08/3_Multinomial.pdf, letzter Aufruf: 24.05.2017

[BEL06] P.P. Belobaba, Airline Demand Analysis and Spill Modeling, 2006, https://ocw.mit.edu/courses/aeronautics-and-astronautics/16-75j-airline-management-spring-2006/lecture-notes/lect4b.pdf, letzter Aufruf: 23.05.2017

[BER51] J. Berkson, Why I prefer logits to probits, Biometrics 7, 1951

[BIL95] P. Billingsley, Probability and Measure, Third Edition, JOHN WILEY & SONS, 1995

[BOE15] Current Market Outlook 2014-2033, Boeing, 2015, http://www.boeing.com/assets/pdf/commercial/cmo/pdf/Boeing_Current_Market_Outlook_2014.pdf, letzter Aufruf: 23.05.2017

[BON08] P.A. Bonnefoy, R.J. Hansman, SCALABILITY OF THE AIR TRANSPORTATION SYSTEM AND DEVELOPMENT OF MULTI-AIRPORT SYSTEMS: A WORLDWIDE PERSPECTIVE, MIT International Center for Air Transportation (ICAT), Report No. ICAT-2008-02 May, 2008

[BRO81] R. Brombacher, A.-W. Scheer, Dezentrales Marketing-Informationssystem - Dialogsystem zur Auswahl geeigneter Datenanalyse- und Progrnoseverfahren, Institut für Wirtschaftsinformatik, Universität des Saarlandes, 1981

[BRO02] M. Brons, E. Pels, P. Nijkamp, p. Rietveld, Price elasticities of demand for passenger air travel: a meta-analysis, Journal of Air Transport Management 8, pp. 165–175, 2002

[BUR03] G. Burghouwt, J. Hakfoort, J.R. van Eck, The spatial configuration of airline networks in Europe, Journal of Air Transport Management 9, pp.309–323, 2003

[BVD] https://de.wikipedia.org/wiki/Bias-Varianz-Dilemma, letzter Aufruf: 22.05.2017

[CKS] http://www.crashkurs-statistik.de/der-korrelationskoeffizient-nach-pearson/, letzter Aufruf: 24.05.2017

[CON04] S.R. Conway, SCALE-FREE NETWORKS AND COMMERCIAL AIR CARRIER TRANSPORTATION IN THE UNITED STATES, 24TH INTERNATIONAL CONGRESS OF THE AERONAUTICAL SCIENCES, 2004,

http://www.icas.org/ICAS_ARCHIVE/ICAS2004/PAPERS/479.PDF, letzter Aufruf: 23.05.2017

[COO08] G.N. Cook, J. Goodwin, Airline Networks: A Comparison of Hub-and-Spoke and Point-to-Point SystemsAirline Networks: A Comparison of Hub-and-Spoke and Point-to-Point Systems, Journal of Aviation/Aerospace Education & Research, Volume 17, Number 2, 2008

[DAG00] J.K. Dagsvik, Probabilistic Models for Qualitative Coice Bahavior - An Introduction, Statistics Norway Research Department, 2000

[DAV03] A.C. Davison, Statistical Models, Cambridge Series in Statstical and Probabilistic Mathematics, CAMBRIDGE UNIVERSITY PRESS, 2003

[DCAA] „Aviation Emissions and Evaluation of Reduction Options (AERO) Main Report“, Director General for Civil Aviation, Ministry of Transport, Public Works and Water Management of the Kingdom of the Netherlands, Dutch Civil Aviation Authority, 2002

[DEU02] P. Deuflhard, A. Hohmann, Numerische Mathematik I - Eine algorithmisch orientierte Einführung, 3. überarbeitete und erweiterte Auflage, de Gruyter, 2002

[DEU04] P. Deuflhard, Newton Methods for Nonlinear Problems - Affine Invariance and Adaptive Algorithms, Springer, 2004

[DIF] http://www.distancesfrom.com, letzter Aufruf: 22.05.2017

[DOY05] J.C. Doyle, D.L. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, W. Willinger, The „robust yet fragile“ nature of the Internet, PNAS, October 11, Vol. 102, no. 41, pp.14497–145022005, 2005

[DLR] http://www.dlr.de/dlr/desktopdefault.aspx/tabid-10443/637_read-251/#/gallery/8570, letzter Aufruf: 22.05.2017

[EFR93] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, CHAPMAN & HALL, 1993

[ELE] http://www.undertec.de/blog/2013/06/fitting-an-elephant—einen-elefanten-anpassen.html, letzter Aufruf: 24.05.2017

[EVA09] A.D. Evans, A. Schäfer, Simulating Flight Routing Network Responses to Airport Capacity Constraints in the US, 9th AIAA Aviation Technology, Integration, and Operations Conference (ATIO), 2009, http://enu.kz/repository/2009/AIAA-2009-6978.pdf, letzter Aufruf: 23.05.2017

[FAA08] Terminal Area Forecast Summary - Fiscal Years 2008-2025, Federal Aviation Administration, https://www.faa.gov/about/office_org/headquarters_offices/apl/aviation_forecasts/taf_reports/media/TAF%20Summary%20Report%20FY%202008-2025.pdf, 2008, letzter Aufruf: 23.05.2017

[FAD73] D. McFadden, Conditional logit analysis of qualitative coice behavior, University of California at Berkeley, 1973, https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf, letzter Aufruf: 23.05.2017

[FAD00] D. McFadden, Disaggregate Behavioral Travel Demand's RUM Side - A 30-Year Retrospective, prepared for presentation at the International Association of Travel Behavior Analysts, Brisbane, Australia, July 2, 2000, https://eml.berkeley.edu/wp/mcfadden0300.pdf, letzter Aufruf: 23.05.2017

[FLE64] R. Fletcher, C. M. Reeves, Function minimization by conjugate gradients, The Computer Journal, 7 (2): 149-154, 1964, http://www.yaroslavvb.com/papers/fletcher-function.pdf, letzter Aufruf: 24.05.2017

[FLS] http://www.flightstats.com, letzter Aufruf: 22.05.2017

[GAB] https://github.com/gabrielelanaro/pyquante/blob/master/PyQuante/optimize.py, letzter Aufruf: 22.05.2017

[GUI05] R. Guimera, S. Mossa, A. Turtschi, L.A.N. Amaral, The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles, PNAS, May 31, Vol. 102, No. 22, pp. 7794–7799, 2005

[HAN09] M. Hanke-Bourgeois, Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens, 3. Auflage, VIEWEG + TEUBNER, 2009

[HAR04] A. Hartmann, Kaufentscheidungsprognose auf Basis von Befragungen: Modelle, Verfahren und Beurteilungskriterien, Gabler Edition Wissenschaft, 2004

[HAS08] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Second Edition, Springer, 2008

[HAU78] J.A. Hausman, D.A. Wise, A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences, Econometrica, Vol26, Nr. 2, Seite 403-426, The Econometric Society, 1978

[HEC06] J.J. Heckman, Probabilistic Choice Models, University of Chicago, Econ 312, 2006

[HOF88] S.D. Hoffmann, G.J. Duncan, Multinomial and Conditional Logit Discrete-Choice Models in Demography, Demograplhy, Vol. 25, No. 3, 1988

[HOL04] B.J. Holmes, Transformations in Air Transportation Systems For the 21st Century - General Lecture, International Council for Aeronautics and Space Twenty-Fourth Congress - Yokohama, Japan, 2004, http://www.icas.org/ICAS_ARCHIVE/ICAS2004/PAPERS/600.PDF, letzter Aufruf: 23.05.2017

[HUJ05] R. Hujer, Folien zur Vorlesung Mikroökonometrie, Johann Wolfgang Goethe-Universität, Frankfurt am Man, 2005, https://www.wiwi.uni-frankfurt.de/professoren/hujer/Lehre/Oek_2/skript.pdf, letzter Aufruf: 23.05.2017

[HWF] http://www.hwf-hamburg.de/wirtschaftsstandort/2036946/luftfahrtindustrie.html, letzter Aufruf: 24.05.2017

[HÖF04] C. Höft, Bewertung von Verfahren zur Prognose der elektrischen Last - eine empirische Analyse, Diplomarbeit, Technische Universität Dresden, DREWAG-Chair for Energy Economics, 2004

[HÜF06] M. Hüftle, Nichtlineare Optimierung ohne Nebenbedingungen, 2006

[IAC] https://de.wikipedia.org/wiki/Liste_der_IATA-Airline-Codes, letzter Aufruf: 22.05.2017

[IBC] http://de.wikipedia.org/wiki/Liste_der_IATA-Bahnhofs-Codes, letzter Aufruf: 22.05.2017

[ICAO04] International Civil Aviation Organization, Outlook for Air Transport to the Year 2015, Approved by the Secretary General and published under his authority, 2004

[ICAO07] International Civil Aviation Organization, Outlook for Air Transport to the Year 2025, Approved by the Secretary General and published under his authority, 2007

[IFC] https://de.wikipedia.org/wiki/Liste_der_IATA-Flughafen-Codes, letzter Aufruf: 22.05.2017

[IGE] C. Igel, M. Hüsken, Empirical Evaluation of the Improved Rprop Learning, Neurocomputing 50, Seite 105-123, 2003

[IMB07] G. Imbens, „Waht's New in Econometrics“ Lecture 11 - Discrete Choice Modells, NBER Summer Institute, 2007

[INN] http://www.innovata-llc.com/about-innovata/, letzter Aufruf: 22.05.2017

[INWT] https://www.inwt-statistics.de/blog-artikel-lesen/Bestimmtheitsmass_R2-Teil2.html, letzter Aufruf: 24.05.2017

[IPDS] http://www.ipds.uni-kiel.de/Dokumente/ModulG/Teil_1/170108_modul_g.pdf, letzter Aufruf: 24.05.2017

[JUD98] G. Judge, R. Carter Hill, Introduction to the Theory and Practice of Econometrics, 1998

[KAL] G. Kallianpur, White Noise Theory of Prediction, Filtering and Smoothing. CRC Press Inc., 1988

[KAM89] W.A. Kamakura, The Estimation of Multinomial Probit Models: A New Calibration Algorithm, Transportation Science, Vol. 23, No. 4, 1989

[KAR75] S. Karlin, H.M. Taylor, A FIRST COURSE IN STOCHASTIC PROCESSES, Second Edition, Elsevier, Academic Press, 1975

[KOELN] http://eswf.uni-koeln.de/glossar/node101.html, letzter Aufruf: 24.05.2017

[KOM] https://de.wikipedia.org/wiki/Komplexität

[KOP06] F.S. Koppelman, C. Bhat, A Self Instructing Course in Mode Choice Modeling: Multinomial and Nested Logit Models, Prepared For U.S. Department of Transportation - Federal Transit Administration, 2006

[KOT09] T. Kotegawa, S. Han, D.A. DeLaurentis, Implementation of Enhanced Network Restructuring Algorithms for Improved Air Traffic Forecasts, Technology, Integration, and Operations Conference, 2009

[KOT10-1] T. Kotegawa, D.A. DeLaurentis, G. Harden, METHODS TO INTEGRATE MULTIPLE STAKEHOLDER PERSPECTIVES INTO AIR TRASNPORTATION EFFICIENCY METRICS, 27TH INTERNATIONAL CONGRESS OF THE AERONAUTICAL SCIENCES, 2010

[KOT10-2] T. Kotegawa, D.A. DeLaurentis, A. Sengstacken, Development of network restructuring models for improved air traffic forecasts, Elsevier, Transportation Research Part C 18, pp. 937–949, 2010

[KOT14] T. Kotegawa, D. Fry, D.A. DeLaurentis, E. Puchaty, Impact of service network topology on air transportation efficiency, Elsevier, Transportation Research Part C 40, pp. 231–250, 2014

[KRI05] D. Kriesel, Ein kleiner Überblick über Neuronale Netze, 2005, www.dkriesel.com, letzter Aufruf: 22.05.2017

[LAN71] K. Lancaster, Consumer Demand: A New Approach,Columbia University Press, 1971

[LAN08] E. Lancsar, J. Louviere, Conducting Discrete Choice Experiments to Inform Healthcare Decision Making, PharmacoEconomics, Band 26, Nr. 8, 2008, Seite 661–677

[LAU06] D.A. DeLaurentis, E. Han, T. Kotegawa, Establishment of a Network-based Simulation of Future Air Transportation Concepts, 6th AIAA Aviation Technology, Integration and Operations Conference, 2006

[LAU08] D.A. DeLaurentis, E. Han, T. Kotegawa, Network-Theoretic Approach for Analyzing Connectivity in Air Transportation Networks, JOURNAL OF AIRCRAFT, Vol. 45, No. 5, 2008

[LEA16] J. Leahy, Global Market Forecast 2016-2035 - Drivers & Results, http://www.airbus.com/company/market/global-market-forecast-2016-2035/, 2016

[LEH14-1] S. Lehner, V. Gollnick, Function-Structure Interdependence of Passenger Air Transportation: Application of a Systemic Approach, 14th AIAA Aviation Technology, Integration, and Operations Conference, 2014

[LEH14-2] S. Lehner, K. Kölker, K. Lütjens, EVALUATING TEMPORAL INTEGRATION OF EUROPEAN AIR TRANSPORT, 29th congress of the International Council of hte Aeronautical Sciences, 2014

[LEH14-3] S. Lehner, V. Gollnick, S. Nishant, Analyzing Collaboration in Self-Organizing Complex Systems of Flow, 2014, http://ieeexplore.ieee.org/document/6819291/, letzter Aufruf: 23.05.2017

[LET] https://www.3lettercode.de, letzter Aufruf: 24.05.2017

[LON] D. Long, D. Lee, J. Johnson, E. Gaier, P. Kostiuk, Modeling Air Traffic Management Technologies With a Queuing Network Model of the National Airspace System, National Aeronautics and Space Administration, Prepared for Langley Research Center under Contract NAS2-14361, 1999,

https://pdfs.semanticscholar.org/b9c8/3126afb91cc91887c02e8d540a07e8063565.pdf, letzter Aufruf: 23.05.2017

[LOU00] J.J. Louviere, D.A. Henscher,J.D. Swait, Stated Choice Methods: Analysis and Application, Cambridge University Press, Cambridge, 2000

[LST] http://www.lernstats.de/php/texte.php?lang=de&sub=korrelation?07_03, letzter Aufruf: 24.05.2017

[MAD83] G. S. Maddala, Limited Dependent and Qualitative Variables in Eco No Metrics, Seite 60-61, 1983

[MAI90] G. Maier, P. Weiss, Modelle diskreter Entscheidungen, Springer-Verlag, Wien, 1990

[MAK06] T. Makowski, Zukunft des Personenluftverkehrs - Flugroutenverkehrs- und -wahlprognose, Nachfrageanalyse durch stochastische Nutzenmaximierung (RUM) sowie szenariotechnische Routen- und Angebotsbestimmung relationsabhängiger Nachfragemengen als Elemente zur Prognose routenspezifischer Passagierzahlen im internationalen Personenluftverkehr, Dissertation der an der Fakultät für Wirtschaftswissenschaften der Rheinisch-Westfälischen Technischen Hochschule Aachen, 2006

[MEY] L.A. Meyn, T.F. Romer, K. Roth, L.J. Bjarke, Preliminary Assessment of Future Operational Concepts Using the Airspace Concept Evaluation System, American Institute of Aeronautics and Astronautics, https://aviationsystemsdivision.arc.nasa.gov/publications/modeling/AIAA-2004-6508.pdf, letzter Aufruf: 24.05.2017

[MUE] R. Müller, Rauschen, 2. Auflage, Springer, 2013

[MUE12] S. Müller, Analyse diskreter Auswahlentscheidungen - Discrete Choice Analysis, 2012, http://docplayer.org/18173336-Analyse-diskreter-auswahlentscheidungen-discrete-choice-analysis.html, letzter Aufruf: 24.05.2017

[NEW] http://www-history.mcs.st-andrews.ac.uk/Biographies/Raphson.html, letzter Aufruf: 22.05.2017

[NEW03] M.E.J. Newman, The Structure and Function of Complex Networks, SIAM REVIEW, Vol. 45,No. 2,pp. 167–256, 2003

[OCC] http://math.ucr.edu/home/baez/physics/General/occam.html, letzter Aufruf: 22.05.2017

[OFS] http://www.openflights.org/, letzter Aufruf: 22.05.2017

[ORT00] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Society for Industrial & Applied Mathematics, 2000

[PAR] http://www.partow.net/miscellaneous/airportdatabase/, letzter Aufruf: 22.05.2017

[PRI] http://data.princeton.edu/wws509/notes/c6.pdf, letzter Aufruf: 22.05.2017

[PRO] http://www.prokerala.com/travel/airports/, letzter Aufruf: 22.05.2017

[RAO73] C. R. Rao, Linear Statistical Inference and Its Applications, 2nd ed. New York: Wiley, 1973

[RED11] R. Redondi, P. Malighetti, S. Paleari, Hub competition and travel times in the world-wide airport network, Journal of Transport Geography 19, pp. 1260–1271, 2011

[REG11] A. Reggiani, P. Nijkamp, A. Cento, Connectivity and Concentration in Airline Networks: A Complexity Analysis of Lufthansa's Network, Tinbergen Institute Discussion Paper, 2011

[REY98] A.J. Reynolds-Feighan, The Impact of U.S. Airline Deregulation on Airport Traffic Patterns, Geographical Analysis, Vol. 30, No. 3, Ohio State University Press, 1998

[REY07] T.G. Reynolds, S. Barrett, L.M. Dray, A.D. Evans, M.O. Köhler, Modelling Environmental & Economic Impacts of Aviation: Introducing the Aviation Integrated Modelling Project, Revised paper for 7th AIAA Aviation Technology, Integration and Operations Conference, Belfast, 18-20 September 2007, Paper No. AIAA-2007-7751, 2007

[RUG] http://www.rug.nl/research/portal/files/9892551/c3.pdf, letzter Aufruf: 23.05.2017

[SAS] ascend, Taking your airline to new heights, A MAGAZINE FOR AIRLINE EXECUTIVES, 2009 Sabre Inc., Ausgabe 2, 2007, https://www.sabreairlinesolutions.com/pdfs/AnalyzeThis_OCT_2007.pdf, letzter Aufruf: 22.05.2017

[SAB] http://www.sabre.com/about/, letzter Aufruf: 22.05.2017

[SBA1] https://www.destatis.de/DE/Methoden/Zeitreihenanalyse.pdf, letzter Aufruf: 22.05.2017

[SBA2] https://www.destatis.de/DE/Publikationen/WirtschaftStatistik/Allgemeines Methoden/UmstellungZeitreihenanalyse111983.pdf, letzter Aufruf: 22.05.2017

[SBA3] https://www.destatis.de/DE/Methoden/Zeitreihen/Software Zeitreihenanalyse.html, letzter Aufruf: 22.05.2017

[SBA15] Benutzerhandbuch zu BV4.1, Version 2.1 (deutschsprachige Programmversion), Statistisches Bundesamt, Juni 2015

[SCH06] C. Schmiedl, Neuronale Netze: mehrschichtige Perzeptrone, Proseminar Machine Learning, 2006

[SES07] A. Seshadri, H. Baik, A.A. Trani, A Model to Estimate Origin-Transfer-Destination Route Flows and Origin-Destination Segment Flows across the Continental United States, TRB 2007 Annual Meeting CD-ROM, 2007, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.583.2000&rep=rep1&type=pdf, letzter Aufruf: 23.05.2017

[SIL] N. Silberhorn, T. Dannewald, Alternative Strukturen des Nested Logit Modells mit der PROC MDC, http://de.saswiki.org/images/9/99/10.KSFE-2006-Silberhorn-Alternative-Strukturen-des-Nested-Logit-Modells-mit-der-PROC-MDC.pdf, ltzter Aufruf: 22.05.2017

[SPE04] H.T. Speth, METHODENBERICHTE, Komponentenzerlegung und Saisonbereinigung ökonomischer Zeitreihen mit dem Verfahren BV4.1, Statistisches Bundesamt, DSTATIS, Heft 3, 2004

[SRM] http://www.control.tu-berlin.de/images/2/24/SRM_Slides_6.pdf, letzter Aufruf: 22.05.2017

[TAY] https://www.math.uni-hamburg.de/teaching/export/tuhh/cm/a2/15/vorlesungen/vl2_dv.pdf, letzter Aufruf: 22.05.2017

[TRA02] K. Train, Discrete Choice Methods with Simulation, Cambridge University Press, 2002

[TSS] https://www.thesius.de/offerings/2545350, letzter Aufruf: 22.05.2017

[TWM] https://tools.wmflabs.org/geohack/, letzter Aufruf: 22.05.2017

[UNS1] https://page.mi.fu-berlin.de/rojas/neural/chapter/K5.pdf, letzter Aufruf: 22.05.2017

[UNS2] https://www.coursera.org/learn/machine-learning/lecture/olRZo/unsupervised-learning, letzter Aufruf: 22.05.2017

[WAC] http://www.world-airport-codes.com/, letzter Aufruf: 22.05.2017

[WAI] I .Waitz, S. Lukachko, K. Willcox, P. Belobaba, E. Garcia, P. Hollingsworth, D. Mavris, K. Harback, F. Morser, M. Steinbach, 2006. „Architecture Study for the Aviation Environmental Portfolio Management Tool“, Partnership for Air Transportation Noise and Emissions Reduction, Report No. PARTNER-COE-2006-002.

[WAN08] L. Wang, G. Donohhue, K. Hoffman, L. Sherry, R. Oseguera-Lohr, Analysis of Air Transportation for the New York Metroplex: Summer 2007, Submitted 2/08 International Conference on Research in Air Transportation (ICRAT 2008), Center for Air Transportation Systems Research (CATSR)/GMU 02/08, 2008, http://www.icrat.org/icrat/seminarContent/2008/Analysis_of_Air_Transportation.pdf, letzter Aufruf: 23.05.2017

[WEB63] E. Weber, C. Brott, Ein Linearitätstest mit Hilfe elektronischer Datenverarbeitungsanlagen, Biometrical Journal, Volume 5, Issue 3, Pages 188–205 , 1963

[WEI05] W. Wei, M. Hansen, Impact of Aircraft Size and Seat Availability on Airlines' Demand and Market Share in Duopoly Markets, Transportation Research Part E: Logistics and Transportation Review, Volume 41, Issue 4, Pages 315–327, 2005

[WIT70] H. Witting, G. Nölle, Angewandte mathematische Statistik, Teubner Verlag, Stuttgart, 1970

[WWW] https://www.wiwiweb.de/statistik/zusammenha/zusammmetri/pearson.html, letzter Aufruf: 24.05.2017

[YAN] T.H. Yang, Air Transport Demand, http://www2.nkfust.edu.tw/ translab/airTrasnport/airTransportDemandEstimation.pdf, letzter Aufruf: 23.05.2017

Erklärung

Hiermit erkläre ich, dass ich meine Arbeit selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die Arbeit noch nicht anderweitig für Prüfungszwecke vorgelegt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Mittweida, 28.07.2017